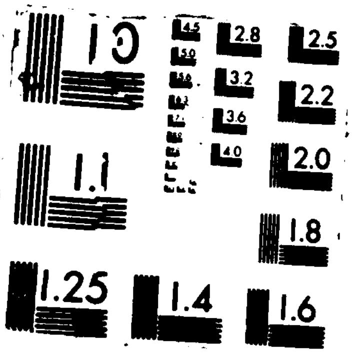
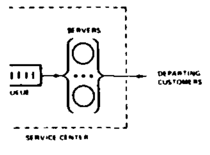
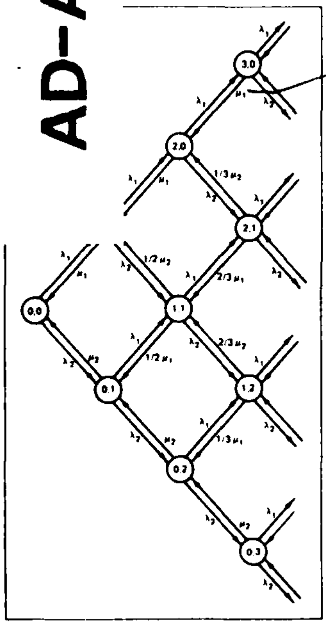
AD-A182 994   NETWORKS OF MARKOVIAN QUEUES(U) ARMY ENGINEER WATERWAYS   1/4
EXPERIMENT STATION VICKSBURG MS INSTRUMENTATION
SERVICES DIV R A FRANCO MAY 87 WES/TR/O-87-2

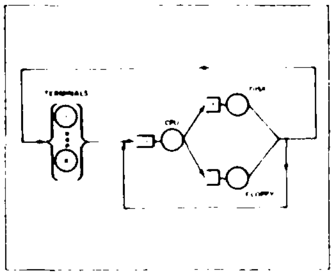
UNCLASSIFIED   F/G 12/3   NL

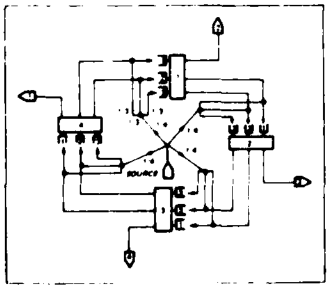

1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

US Army Corps
of Engineers









(2)

TECHNICAL REPORT O-87-2

NETWORKS OF MARKOVIAN QUEUES

by

Raphael A. Franco, Jr.

Instrumentation Services Division

DEPARTMENT OF THE ARMY
Waterways Experiment Station, Corps of Engineers
PO Box 631, Vicksburg, Mississippi 39180-0631

May 1987
Final Report

Approved For Public Release; Distribution Unlimited

Prepared for DEPARTMENT OF THE ARMY
US Army Corps of Engineers
Washington, DC 20314-1000

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report 0-87-2	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION USAEWES, Instrumentation Services Division 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) PO Box 631 Vicksburg, MS 39180-0631	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Networks of Markovian Queues

12. PERSONAL AUTHOR(S)
Franco, Raphael A., Jr.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final report	FROM ______ TO ______	May 1987	300

16. SUPPLEMENTARY NOTATION
Available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Markovian processes (LC) Queueing theory, (LC) Stochastic processes (LC)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

There is a recognized need to make the subject of queueing network theory less esoteric. The engineer who is faced with an application often does not understand the concepts. The likely result is that he either cannot apply them at all or does so invalidly. In any event, he is faced with a formidable research task. This work is intended to make the task less difficult and the subject less esoteric. Much of the thrust of this text was to examine queueing network theory as presented in the literature, to reinforce the results by independent justifications, reduce the ambiguity resident in some explanations, and present numerous corroborating examples where none were found. The result is that the existing literature, in the area covered, has been expanded in explanation, critiqued as to usage, and delineated as to limitations.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL

PREFACE

This report was originally submitted to Mississippi State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Electrical Engineering.

This report was written by Raphael A. Franco, Jr., of the Instrumentation Services Division (ISD), US Army Engineer Waterways Experiment Station (WES), under the direct supervision of Mr. George P. Bonner, Chief, ISD.

Many people contributed to the completion of this work, both directly and indirectly. First and foremost the author would like to thank Dr. John K. Owens for his encouragement, for without it, it is doubtful that the work would have been completed. Credit for the instigation of the research belongs to Dr. Frank M. Ingels, who provided advice and encouragement throughout the effort. The author would also like to express his gratitude to Mr. Bonner for allowing time away from his job in order to finish the project. In addition, the author wishes to express thanks to the National Aeronautics and Space Administration for partially funding this work. And last, but not least, special thanks go to Mr. Nick Lavecchia for proofreading this document and helping to get it into final form.

Commander and Director of WES was COL Dwayne G. Lee, CE. Technical Director was Dr. Robert W. Whalin.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF FIGURES (CONTINUED)

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO QUEUEING THEORY

1.1 Introduction

A queue is a waiting line of customers at a service center. Depending on the service center, service is provided by either a single server or multiple servers which operate in parallel. Figure 1.1 depicts a typical service center. Customers arrive at the service center, wait in line for a server to become free, receive service, and depart. For a service center to be stable, the mean demand for service cannot be greater than the capacity that can be provided.

However, spurious arrivals and statistical fluctuations in service requirements can temporally cause demand to exceed the capacity. When this occurs, a queue of waiting customers will form.

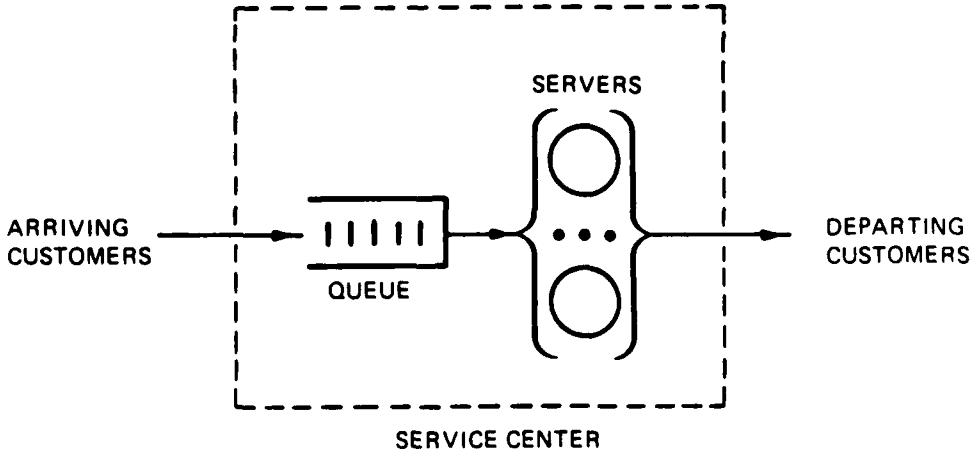


Figure 1.1 A Typical Service Center.

The purpose of a queueing model is to predict the performance of a physical system in which there is contention for resources. The resources are represented in the model by the servers. Any realistic queueing model must incorporate statistical parameters. For example, it is usually not possible to predict with certainty when the next customer will arrive or what his exact service time will be. However, it is often possible to assign probabilities to possible values or continuous intervals of possible values. That is, these parameters are usually random variables that can best be characterized by probability distribution functions. Since the parameters that describe a queueing model are random variables, the performance parameters are also random variables. Thus, queueing models are used to answer probabilistic questions such as: what is the probability that a service center will have k customers at time t, or what is the probability that the waiting time of an arbitrary customer is less than some fixed value x?

Probably the best way to illustrate a queueing model and demonstrate its purpose is by example. Figure 1.2 is a model of a small batch computer system. The labels on the arcs are routing probabilities. It is assumed that the critical resources are the central processing unit (CPU), a hard disk, and a floppy disk. These are represented in the model by single server service centers. A customer or job arrives from outside the system and waits in a queue at the CPU for service. After receiving some service at the CPU, the job requires a hard or floppy disk operation before it can proceed (the probability that it needs the hard disk is 0.8, and the probability it

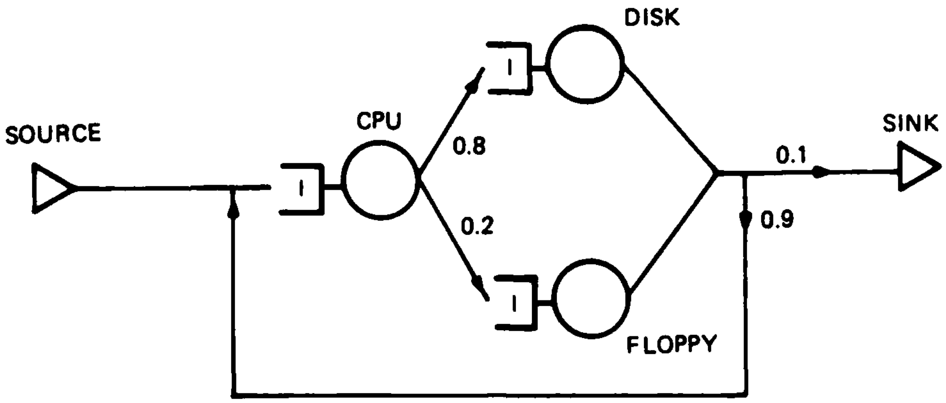


Figure 1.2 Example of an Open Queueing Network Model.

needs the floppy disk is 0.2). Since these operations are usually slow compared to those at the CPU, the CPU releases the job and starts to work on another one. The released job proceeds to the queue at the appropriate disk and waits for service. After receiving disk service the job either exits the system or return to the queue at the CPU for more service (the probability it exits the system is 0.1, and the probability it returns to the CPU is 0.9).

In order to analyze such a model the arrival process, the routing probabilities, the service demand, and the order in which jobs are served must be described. Some of the performance parameters that can be obtained from the model are : the mean waiting time, the mean response, the mean throughput, the mean number of jobs at each service center, and the utilization at each service center.

The model is Figure 1.2 is classified as an open queueing network model. The word network refers to the fact that there is more than one service center in the model. The word open is in reference to the arrival process. In an open queueing model customers arrive from outside of the system and leave the system once their service demand has been met.

Figure 1.3 depicts a closed queueing network model of an interactive computer system. The structure of the model is the same as

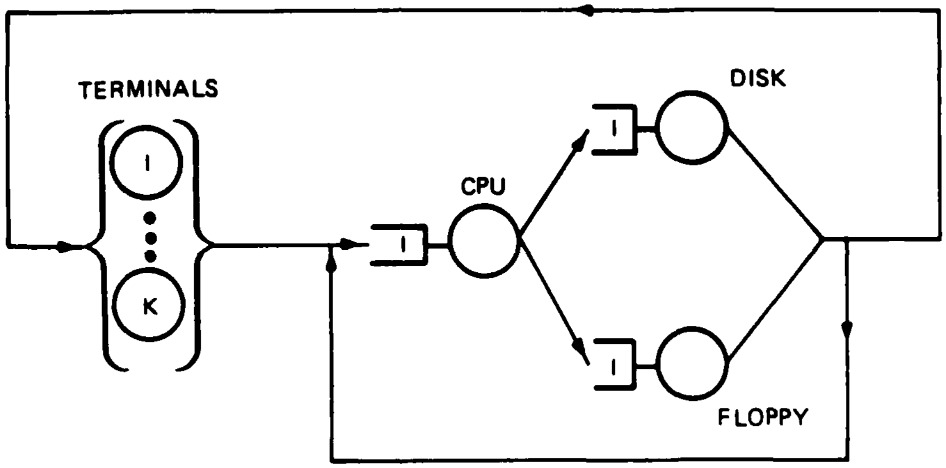


Figure 1.3 Example of Closed Queueing Network Model.

the one in Figure 1.2 with the exception of how customers enter the system. In this model there are a finite number of terminals, K, and it is assumed that there is always one customer at each terminal. Thus, once a customer's service demand is met, he is immediately replaced by a new customer. Clearly, the system is equivalent to one in which customers can neither enter or leave the system, and hence the name closed. In addition, the system is self regulating in that is impossible for the mean service demand to exceed the capacity. If a new customer tries to enter the system when it contains K customers, he is simple turned away. Although the structure of this model is similar to the one in Figure 1.2 the service demand and routing probabilities are usually quite different. In addition, the parameter that one usually varies in an open network is the arrival rate, whereas in a closed network it is K.

The focus of this text is queueing network theory, however before one can analyze a network of service centers, he must first learn how to analyze single service centers. The rest of the chapter and several more are devoted to this.

1.2 Steady-State and Equilibrium

Since a queueing model of a physical system is a probability model, the number of customers in a service center at time t is a discrete random variable. Let N(t) denote this random variable and let $P_k(t) = P[N(t)=k]$. That is, $P_k(t)$ is the probability of finding k customers in the system at time t. $P_k(t)$ depends not only on the value of t, but also on the number of customers in the center at t=0. For small values of t, the number of customers in the service center will be largely influenced by the number of initial customers. However, as t becomes larger the influence will become less, and after a sufficient period of time the number of customers in the service center will become effectively independent of the initial number of customers. The situation is very similar to those found in electrical circuits that contain inductors and/or capacitors, and the same terminology is used. The time dependent solution of $P_k(t)$ is called the transient solution, and the time independent solution the steady-state solution.

The steady-state solution is denoted P_k and defined to be

$$P_k = \lim_{t \to \infty} P_k(t) \ . \qquad (1.1)$$

P_k is the long-term probability of finding exactly k customers in the service center. It is important to understand that whereas P_k is no longer an explicit function of t, the number of customers in a service center will certainly change with time. That is P_k equals the long-term proportion of time that the service center contains exactly k customers. P_k is also often referred to as the equilibrium solution because ultimately the flow of customers into a service center must equal the flow of customers out of the service center.

Queueing theory focuses primarily on the steady-state solution. This is not only because of the difficulties in obtaining transient solutions, but also because the extra information contained in them is of little use.

1.3 Specification of Queueing Systems

As previously mentioned, before a service center can be analyzed it is first necessary to specify the stochastic processes that describe the arriving customers, and the structure and discipline of the service center. Generally the arrival process is described in terms of the probability distribution of the interarrival times (that is the times between successive arrivals of customers at the service center). The assumptions used in most of queueing theory are that these interarrival times are both independent and identically distributed random variables. Hence, they all have the same probability distribution function (PDF) which describes the arrival process. The arrival process is denoted by A(t) and is by definition:

$$A(t) = P[\text{time between arrivals} \le t]. \tag{1.2}$$

In order to satisfy the assumptions of independent and identically distributed random variables, it may be necessary to partition the arrival stream into classes of customers and define a PDF for each class.

The second statistical quantity that must be described is the service time. It is usually assumed that the service times are independent random variables all having the same PDF. The service time PDF is denoted by $B(\tau)$ and defined to be:

$$B(\tau) = P[\text{service time} \le \tau]. \tag{1.3}$$

Again it may be necessary to partition the customers into classes and define a service time distribution for each class.

One must specify a variety of additional quantities in order to identify the structure and discipline of the service center. The first of these is the number of parallel servers at the service center. Another is the available storage capacity to hold additional customers. Often this quantity is assumed to be infinite. Still another specification is the customer population. That is the maximum number of customers that can simultaneously require service. Again this quantity is often assumed to be infinity. In addition to these it may be necessary to specify the service discipline or order in which customers are served. Typical service disciplines are first-come first-served (FCFS), last-come first-served (LCFS), processor sharing (round-robin), and priority.

1.4 Shorthand Notation for Queueing Systems

The shorthand notation A/B/m/K/M is commonly used to describe a queueing system consisting of a single service center. Here A describes the interarrival time distribution, B the service time distribution, m the number of parallel servers, K the system's storage capacity, and M – the customer population. If any of the descriptors are absent, then it is assumed that they take the value of infinity. Thus, if it is assumed that the storage capacity and customer population are infinite, only the first three descriptors are required. The following is a list of well-accepted symbols for distributions:

M	Exponential distribution (i.e., Markovian)
D	Deterministic variable, a constant
E_k	k-stage Erlangian distribution
H_k	k-stage hyperexponential distribution
G	General distribution.

For example, the notation M/D/1 implies a single server system with exponential arrivals and a constant (deterministic) service time.

1.5 Little's Law

Probably the simplest, yet the most important formula, in queueing theory is Little's law [LITT61]. It states that the mean response time R, is equal to the mean queue length L, divided by the mean arrival rate λ. That is,

$$R = \frac{L}{\lambda} \tag{1.4}$$

To show that Little's law is valid, consider Figure 1.4 which is a plot of the number of customers in a service center versus time.

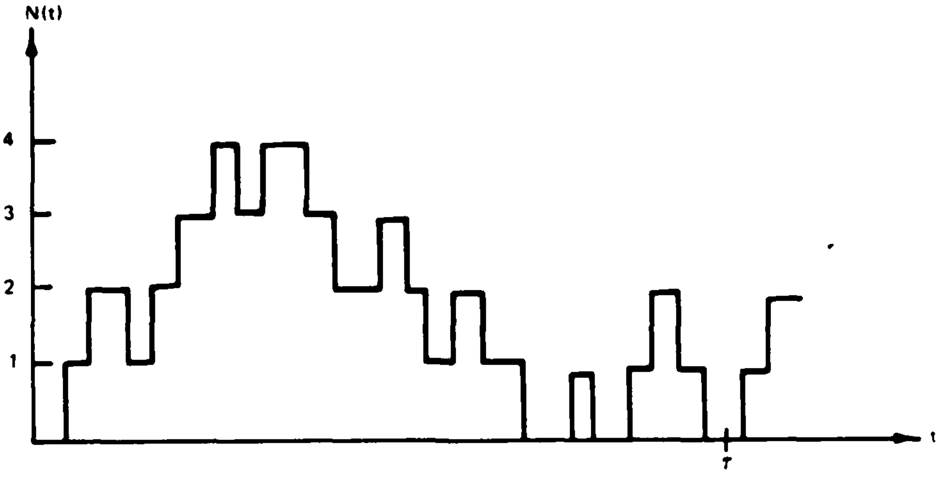


Figure 1.4 Plot of Customers versus Time in a Typical Service Center.

Let

N(t) = the number of customers in the system at time t,

α(τ) = the number of customers that arrive in the interval [0,τ],

ξ(τ) = the area under the curve in interval [0,τ].

During the interval $[0,\tau]$ the mean arrival rate is

$$\lambda = \frac{\alpha(\tau)}{\tau} , \tag{1.5}$$

and the mean number of customers in the system is

$$L = \frac{1}{\tau} \int_0^\tau N(t)\ dt = \frac{\xi(\tau)}{\tau} . \tag{1.6}$$

The area under the curve during this interval equals the total number of customer-seconds spent in the system by the $\alpha(\tau)$ customers. If the number of customers in the system at τ equals the the number of initial customers, then the mean time spent in the system per customer is

$$R = \frac{\xi(\tau)}{\alpha(\tau)} . \tag{1.7}$$

Hence,

$$\frac{L}{\lambda} = \frac{\xi(\tau)}{\tau} \frac{\tau}{\alpha(\tau)} = \frac{\xi(\tau)}{\alpha(\tau)} = R . \tag{1.8}$$

The stipulation that the interval be chosen such that the number of initial and final customers in the system be equal, is nothing more than the steady-state or equilibrium condition. That is, over the long-run the number of customers that flow into a system must equal the number of customers that flow out of the system. This implies that the throughput equals the arrival rate. Thus Little's law can be also be expressed as

$$R = \frac{L}{T} , \tag{1.9}$$

where T represents the throughput. There are no standard notations in queueing theory, hence different authors use different symbols to represent the quantities in Little's law. Although the same symbols often appear in usage by different authors, they are used to symbolize different quanties. In order to avoid confusion, the second expression will almost always be used in this text since the letters have intuitive meaning.

It is important to emphasize that Little's law does not depend on any specific assumptions regarding the arrival or service time distributions, nor does it depend on the number of servers or the order in which customers are served. It holds for any system in which customers arrive, wait for service, and depart. It does not matter if the system is composed of a single service center or a collection of service centers. In fact Little's law can even be applied to parts of a service center. For example, if L_q is the number of customers waiting to be served and W_q the mean waiting time, then $L_q = \lambda W_q$.

1.6 Utilization

The utilization of a service center is the average amount of service required divided by the maximum amount of service that can be provided. If the arrival and service processes are independent of each other and of the number of customers in the system, then on the average λ customers arrive per second, and each customer requires E[S] seconds of service. Thus the average amount of service required per second is $\lambda E[S]$. Now for a single-server system, the maximum amount of service that can be provided is one second of service per second. Hence, the utilization (denoted by ρ) of a single-server system is

$$\rho = \lambda E[S] \ . \tag{1.10}$$

For a service center with m servers the maximum amount of service that can be provided is m seconds of service per second. Therefore, the utilization for a multi-server system is

$$\rho = \frac{\lambda E[S]}{m} \ . \tag{1.11}$$

Clearly, for a single-server system, the utilization is the proportion of time the server is busy. Similarly, for a multi-server system, utilization is the average proportion of time the servers are busy. Since P_k is the long-run proportion of time the system contains k customers, P_0 is the long-run proportion of the time that the system is empty or not busy. Now since the summation of the P_k's over all k must equal one, the proportion of time that the system is busy is $1-P_0$. Thus for a single-server system utilization can be expressed as

$$\rho = 1-P_0 \ . \tag{1.12}$$

Similar results can be obtained for a multi-server service center. More precisely,

$$\rho = 1 - P_0 - \frac{1}{m}\sum_{k=1}^{m-1}(m-k)\ P_k \ . \tag{1.13}$$

This follows from the fact that when the service center contains k customers (m-k)/m is the capacity of the service center that is not utilized. The last two equation for ρ are more general than the first two, in that the arrival rate does not appear explicitly and, therefore, it may be a function of the number of customers in the system.

Obviously, the utilization of a service center must be less than one. That is, for the system to be stable, the mean service demand cannot exceed the capacity of the system to provide service. If the system is not stable, then as t approaches infinity, the queue length grows without bound and the limiting probabilities do not exist.

1.7 Outline of Contents

The purpose of this short chapter was to introduce queueing theory and some of the terminology that will be used in the following chapters. The next chapter is a mathematical treatment of Markov processes. The tools developed in this chapter form the basis of queueing theory analysis. Chapter 3 is the analysis of Markovian queues, M/M/m. Chapter 4 is the analysis of semi-Markovian queues, M/G/1 and G/M/1. Chapter 5 is an introduction to queueing network theory (Jackson type networks). Chapter 6 is advanced queueing network theory. Chapter 7 is computation algorithms for closed and mixed networks. Finally, Chapter 8 points out the limitations of queueing theory and open areas of research.

CHAPTER 2

MARKOV PROCESSES

2.1 Random Variables

A random variable, X, is a variable whose value depends on the outcome of a random experiment. The outcome of the experiment assigns a value to X. The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S. Each outcome 's' in the set S is referred to as a sample point. Thus a random variable is nothing more than a function defined on the sample space of a random experiment. Therefore, the symbol for a random variable should be $\{X(s):s \epsilon S\}$, denoting the dependence of X on the sample space, but it is customary to use the short hand notation X.

2.2 Stochastic Processes

A stochastic process, $\{X(t,s):t \epsilon T, s \epsilon S\}$ is a family of random variables that describes the evolution through time of some process. The symbol, $\{X(t,s):t \epsilon T, s \epsilon S\}$, indicates that it is a set-function of two variables. The set T represents time, and is often referred to as the index set since for each $t \epsilon T$ (a specific value of t), X(t,s) reduces to a random variable (note that in this chapter and only in this chapter, the variable T will represent time and not throughput). Thus, the variable t induces a set or family of random variables. The set S represents the sample space of these random variables, and s is a sample point in S. Just as it is customary to use the symbol X for a random variable rather than $\{X(s):s\epsilon S\}$, it is also customary to drop the s in the notation of a stochastic process. That is, the traditional symbol for a stochastic process is $\{X(t):t\epsilon T\}$ [PARZ62], [THOM69], [ROSS80].

There are two other notations that are used to denote special stochastic processes. If the set T is finite or countable, then the process is said to be a discrete-time stochastic process. A countable set is one in which there exists a one-to-one correspondence between each element of the set and the nonnegative integers. Therefore, a discrete-time stochastic process is a sequence of random variables indexed by the set T. When the sequence is infinite but countable, the process is often represented by $\{X_n,n=0,1,2,\ldots\}$. If on the other hand, the set T consists of all points on a continuous interval of the time axis, the process is called a continuous-time stochastic process. If the interval consists of the entire positive time axis, then frequently the short hand notation X(t) is used to represent the process.

Stochastic processes are also classified according to the set S. The set S is called the state space, and it consists of all possible values (states) that the random variables may assume. If S is finite or countable, the process is said to be a discrete-state process. Otherwise it is said to be a continuous-state process.

If $X_n=i$, then the process is said to be in state i at time n, or for the continuous-time case if $X(t)=i$ the process is in state i at time t.

2.3 Introduction to Markov Processes

A Markov process is a stochastic process that has no 'memory'. This means that information about how the process reached a certain state plays no roll in assigning probabilities to the next or future states. Only discrete-state Markov processes will be discussed in this text.

2.4 Discrete-State, Discrete-Time, Markov Processes

A discrete-state, discrete-time, Markov process is a stochastic process $\{X_n, n=0,1,2,...\}$ such that:

$$P[X_{n+1}=j \mid X_n=i, X_{n-1}=i_{n-1}, \ldots, X_1=i_1, X_0=i_0] = P[X_{n+1}=j \mid X_n=i_n] \qquad (2.1)$$

for all states $i_0, i_1, \ldots, i_{n-1}, i, j$, and all $n \geq 0$. In other words, the probability of any future state X_{n+1}, given present state X_n and the past states $X_{n-1}, \ldots, X_1, X_0$, depends only on the present state and is independent of the past states. Discrete-time Markov processes are often referred to as Markov chains.

If the probabilities are time-invariant, that is independent of n, then the process is referred to as a homogeneous Markov process. For a homogeneous Markov process, let P_{ij} denote the probability of going from state i to j in one step. That is,

$$P_{ij} = P[X_{n+1}=j \mid X_n=i] \; . \qquad (2.2)$$

The probability of going from any state to another state in one step can be described by the matrix

$$[P] = \begin{vmatrix} P_{00} & P_{01} & P_{02} & \cdots & P_{0j} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots & P_{1j} & \cdots \\ \cdot & \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & & \cdot & \\ P_{i0} & P_{i1} & P_{i2} & \cdots & P_{ij} & \cdots \\ \cdot & \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & & \cdot & \end{vmatrix} \qquad (2.3)$$

where the row index is the present state and the column index the next state. The matrix [P] is called the one-step transition matrix.

Let $P_{ij}{}^2$ denote the probability that a process goes from state i to state j in two steps. In order to go from state i to state j in two steps, the process must go to some intermediate state k in the first step and proceed to state j in the next step. Therefore,

$$P_{ij}{}^2 = \sum_{k=0}^{\infty} P_{ik} P_{kj}$$

$$= P_{i0} P_{0j} + P_{i1} P_{1j} + P_{i2} P_{2j} + \cdots \quad . \qquad (2.4)$$

This equation can be interpreted as the weighted average of going to state j in one step, given the process was in state k, weighted by the probability of being in state k. The two-step transition matrix, denoted by $[P^2]$, can be found from the one-step matrix [P] by observing that the two-step transition $P_{ij}{}^2$ is the sum of the elements along the ith row multiplied by the elements along the jth column. Hence, the two-step transition matrix is

$$[P^2] = [P][P] = [P]^2 \ . \qquad (2.5)$$

It follows that for a process to go from state i to j in n+m steps, it must go to some intermediate state k in n steps and proceed to state j in m steps. Therefore,

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m} \quad . \tag{2.6}$$

The last equation is called the Chapman-Kolmogorov equation. When n=1, the equation is referred to as the backward Chapman-Kolmogorov equation, since it is written at the backwards most end of the interval. More precisely, the backward Chapman-Kolmogorov equation is

$$P_{ij}^{1+m} = \sum_{k=0}^{\infty} P_{ik} P_{kj}^{m}$$

$$= P_{i0} P_{0j}^{m} + P_{i1} P_{1j}^{m} + P_{i2} P_{2j}^{m} + \cdots \quad . \tag{2.7}$$

Note that P_{ij}^{1+m} is the sum of the elements, along the ith row of the one-step transition matrix, multiplied by the elements along the jth column of the m-step matrix. Hence,

$$[P^{1+m}] = [P]\,[P^{m}] \quad . \tag{2.8}$$

If m equals two, then the three-step transition matrix is just the one-step matrix times the two-step matrix, which is the one-step matrix raised to the third power. Recursively it follows that the nth-step matrix is just the one-step matrix raised to the nth power. That is,

$$[P^{n}] = [P]^{n} \quad . \tag{2.9}$$

The same results can be derived by letting m=1. More precisely,

$$P_{ij}^{n+1} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}$$

$$= P_{i0}^{n} P_{0j} + P_{i1}^{n} P_{1j} + P_{i2}^{n} P_{2j} + \cdots \quad , \tag{2.10}$$

or in matrix form

$$[P^{n+1}] = [P^n]\ [P]\ . \qquad (2.11)$$

When m=1 the Chapman-Kolmogorov equation is referred to as the forward Chapmam-Kolmogorov equation, since it is written at the forward most end of the interval.

The unconditional probability of being in state j after n steps, denoted by $P_j{}^n$, is the weighted average of going to state j in n steps, given the initial state was k, weighted by the probability of state k being the initial state. Therefore,

$$P_j{}^n = \sum_{k=0}^{\infty} P_k{}^0\ P_{kj}{}^n$$

$$=P_0{}^0\ P_{0j}{}^n + P_1{}^0\ P_{1j}{}^n + P_2{}^0\ P_{2j}{}^n + \cdots \qquad (2.12)$$

where $P_k{}^0$ is the probability of initially being in state k. Let V^0 equal the vector of initial state probabilities, and let V^n equal the vector of state probabilities after n-steps. The unconditional probabilities in vector form are

$$V^n = V^0\ [P^n] \qquad (2.13)$$

where $V^0 = [P_0{}^0, P_1{}^0, P_2{}^0, \cdots]$ and $V^n = [P_0{}^n, P_1{}^n, P_2{}^n, \cdots]$. Thus, given an initial state probability vector and the one-step transition matrix, one can find the n-step probability vector, that is, the probabilities of where the process will be n steps after start-up.

2.4.1 Limiting Probabilities

When a process first starts up, its initial state certainly has a large influence on the current state, but what about after the process has been operating for a long time? It seems reasonable to expect that as time increases the influence of the initial state should decrease. More specifically, does the limit of P_j^n as n approaches infinity, converge to some value, say P_j, which is independent of the initial state, and do the P_j's form a probability system? The answers to these questions are that it depends on the process.

In order to define when the limiting probabilities exist, it is first necessary to discuss some terminology of Markov chains. A Markov chain is said to be irreducible if every state can be reached from every other state. More precisely, for each pair of states (i and j) there must exist an integer m (which may depend upon i and j) such that $P_{ij}^m > 0$. Furthermore, state i is said to have period n if, when in state i, the only possible steps at which the process can return to state i are n,2n,3n,....,. If n = 1 then state i is aperiodic. It can be shown that all states of an irreducible Markov chain have the same period [ROSS80].

The following theorem states when the limiting probabilities exist: (1) In an irreducible, aperiodic, homogeneous, Markov chain the limiting probabilities

$$P_j = \lim_{n \to \infty} P_j^n = \lim_{n \to \infty} P_{ij}^n \qquad (2.14)$$

always exist and are independent of the initial state probabilities. (2) If the chain is finite, then the limiting probabilities form a probability system. That is,

$$0 \leq P_j \leq 1 \qquad \text{and} \qquad \sum P_j = 1 \quad \text{for all j.} \qquad (2.15)$$

(3) If the chain is infinite, the limiting probabilities may or may not form a probability system, but if they do not then $P_j = 0$ for all j. The interested reader is referred to [FELL66] for a proof.

Assuming that the limiting probabilities do exist, it is easy to show that they are independent of the initial state. The unconditional probability that the process is in state j at step n can be expressed as

$$P_j^n = \sum_{k=0}^{\infty} P_k^{n-1} P_{kj} \quad , \qquad (2.16)$$

or in vector form

$$V^n = V^{n-1} [P] \quad . \qquad (2.17)$$

By taking the limit of these equations as n approaches infinity , one arrives at

$$P_j = \sum_{k=0}^{\infty} P_k P_{kj} \qquad (2.18)$$

or in vector form

$$V = V [P] \qquad (2.19)$$

where $V = [P_0, P_1, P_2, \cdots]$. Therefore, the limiting probabilities are independent of the initial state of the process. Equation (2.18) or (2.19) along with the conservation of probability equation (summation of P_j's equal one) uniquely determines the limiting probabilities when they exist.

It can also be shown that when the limiting probabilities exist then

$$P_j = \lim_{n \to \infty} P_{ij}^n \quad . \qquad (2.20)$$

The equivalent statement in matrix form is that as n approaches infinity the n-step transition matrix approaches a matrix in which each row approaches the vector V. That is,

$$\lim_{n\to\infty} [P^n] = [V] \quad . \tag{2.21}$$

The limit of Equation (2.13) as n approaches infinity is

$$V = V^0 \lim_{n\to\infty} [P^n] \ . \tag{2.22}$$

The initial state vector, V^0, appears explicitly in this equation, and since it is always true that the elements in V^0 form a probability system, then the only way that the equation can hold and be independent of V^0 is if

$$\lim_{n\to\infty} [P^n] = [V]. \tag{2.23}$$

The limiting probabilities are also called the steady-state probabilities, since they represent the states of the process after the *effects of the initial conditions have died out.* It is important to understand that as n approaches infinity the process still moves from state to state, and hence the limiting probabilities equal the long-run proportion of time the process spends in each state.

2.5 Discrete-State, Continuous-Time, Markov Processes

A discrete-state, continuous-time Markov process is a stochastic process $\{X(t), t\geq 0\}$ such that for all $s,t\geq 0$ and nonnegative integers $i,j,x(u),0\leq u<s$

$$P[X(t+s)=j\,|\,X(s)=i, X(u)=x(u), 0\leq u<s] = P[X(t+s)=j\,|\,X(s)=i] \ . \tag{2.24}$$

That is, the probability of the future $X(t+s)=j$ given the present $X(s)$ and past $X(u), 0\leq u<s$, depends only on the present and is independent of

the past. If in addition,

$$P[X(t+s)=j|X(s)=i] \tag{2.25}$$

is independent of s, then the process is said to be stationary or homogeneous. Only homogeneous Markov processes will be considered in this text.

Since the past history includes how long the process has been in current state, the definition requires that the amount of time in the current state and the next state visited must be independent random variables. The definition also requires that if T_i is the random variable representing the amount of time in state i, then the probability distribution function (PDF) of T_i must be 'memoryless'. In other words, the amount of time in state i cannot affect the probability of when the the process will depart state i. The memoryless statement in mathematical terms is

$$P[T_i>t+s|T_i>s] = P[T_i>t] \ . \tag{2.26}$$

The only PDF which has this property is the negative exponential,

$$F_i(t) = P[T_i \leq t] = 1-e^{-\xi_i t} \tag{2.27}$$

where $1/\xi_i$ is the expected value of T_i. The subscripts indicate that T_i may depend on the state i.

The following shows that this distribution function has the memoryless property. The condition in Equation (2.26) is equivalent to

$$P[T_i>t+s|T_i>s] \ P[T_i>s] = P[T_i>t] \ P[T_i>s] \tag{2.28}$$

or

$$P[T_i>t+s] = P[T_i>t] \ P[T_i>s] \ . \tag{2.29}$$

Since

$$e^{-\xi_i(t+s)} = e^{-\xi_i t} \; e^{-\xi_i s} \; , \tag{2.30}$$

it follows that the negative exponential satisfies the condition in Equation (2.29), and therefore has the memoryless property.

The following proves that it is the only distribution function with the memoryless property. The derivative of Equation (2.29) with respect to s is

$$\frac{dP[T_i>t+s]}{ds} = P[T_i>t] \, \frac{dP[T_i>s]}{ds} \; . \tag{2.31}$$

For any PDF

$$\frac{dP[T_i>s]}{ds} = \frac{d\{1-P[T_i\leq s]\}}{ds} = -f_i(s) \; , \tag{2.32}$$

where $f_i(s)$ is the probability density function (pdf). Substituting (2.32) into (2.31) yields

$$\frac{dP[T_i>t+s]}{ds} = -f_i(s) \; P[T_i>t] \; . \tag{2.33}$$

By dividing both sides of this equation by $P[T_i>t]$ and letting s=0, one obtains

$$\frac{dP[T_i>t]}{P[T_i>t]} = -f_i(0) \; ds \; . \tag{2.34}$$

It follows by integrating from 0 to t that

$$\text{Log}_2 \; P[T_i>t] = -f_i(0)t \; , \tag{2.35}$$

$$P[T_i>t] = e^{-f_i(0)t} , \tag{2.36}$$

$$P[T_i\leq t] = 1- e^{-f_i(0)t} . \tag{2.37}$$

Thus, the negative exponential is the only PDF with the 'memoryless' property.

This connection is so strong that the definition of a homogeneous continuous-time Markov process can be given in terms of it. Namely, it is a stochastic process having the properties that each time it enters state i :

(1) the amount of time it spends in state i is exponentially distributed with mean $1/\xi_i$, and

(2) when the process departs state i, it enters state j according to probability p_{ij}. Of course, the p_{ij} must satisfy

$$p_{ii} = 0 \quad \text{for all } i$$
$$\sum_j p_{ij} = 1 \quad \text{for all } i \,. \tag{2.38}$$

For a homogeneous, continuous-time Markov process, let $P_{ij}(t)$ denote the probability of going from state i to state j in time t. More precisely,

$$P_{ij}(t) = P[X(t+s)=j \,|\, X(s)=i] \quad \text{for } s,t \geq 0 \,. \tag{2.39}$$

$P_{ij}(t)$ is analogous to the discrete-time n-step transition probability $P_{ij}{}^n$. The difference is that the discrete parameter n has been replaced by the continuous parameter t. In other words, $P_{ij}{}^n$ is the probability of going from state i to state j in n-steps, and $P_{ij}(t)$ is the probability of going from state i to state j in time t.

Since by definition $P_{ij}(t)$ is independent of s, then $P_{ij}(t)$ is also the conditional probability that the process is in state j at time t, given that it was initially in state i at time 0. That is,

$$P_{ij}(t) = P[X(t)=j \,|\, X(0)=i] \,. \tag{2.40}$$

Let $P_j(t)$ denote the unconditional probability that the process is in state j at time t. More precisely,

$$P_j(t) = \sum_{i=0}^{\infty} P[X(t)=j \mid X(0)=i]\ P[X(0)=i] \ . \qquad (2.41)$$

It follows that

$$\begin{aligned} P_j(t) &= \sum_{i=0}^{\infty} P_{ij}(t)\ P_i(0) \\ &= P_0(0)P_{0j}(t) + P_1(0)P_{1j}(t) + P_2(0)P_{2j}(t) + \cdots \ . \qquad (2.42) \end{aligned}$$

As in the case of the discrete-time Markov process, these probabilities can be expressed in vector and matrix form. That is,

$$V(t) = V(0)\ [P(t)]$$

where $V(t) = [P_0(t), P_1(t), P_2(t), \cdots]$,

and $V(0) = [P_0(0), P_1(0), P_2(0), \cdots]$,

and

$$[P(t)] = \begin{vmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) & \dots & P_{0j}(t) & \dots \\ P_{10}(t) & P_{11}(t) & P_{12}(t) & \dots & P_{1j}(t) & \dots \\ \vdots & \vdots & \vdots & & \vdots & \\ P_{i0}(t) & P_{i1}(t) & P_{i2}(t) & \dots & P_{ij}(t) & \dots \\ \vdots & \vdots & \vdots & & \vdots & \end{vmatrix} \ . \qquad (2.43)$$

Hence given V(0) and [P(t)], the unconditional probabilities of where the process is at time t, V(t), can be computed. The only remaining problem is determining [P(t)].

In the discrete-time Markov chain $P_{ij}{}^{n}$ was derived from the

Chapmam-Kolmogorov equation,

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m} , \qquad (2.6)$$

and a similar procedure will be used here. In order for a continuous-time Markov process to make a transition from state i to state j in time h+t, it must go to some state k in time h and then proceed to state j in the remaining time t. Therefore,

$$P_{ij}(h+t) = \sum_{k=0}^{\infty} P_{ik}(h) \; P_{kj}(t) \; . \qquad (2.44)$$

This equation is the continuous-time equivalent of the Chapmam-Kolmogorov equation. By writing out the k=i term, subtracting $P_{ij}(t)$ from both sides, dividing by h and taking the limit as h approaches zero, one arrives at

$$\lim_{h\to 0} \left[\frac{P_{ij}(h+t) - P_{ij}(t)}{h}\right] = \lim_{h\to 0} \left[\frac{P_{ii}(h)P_{ij}(t) - P_{ij}(t)}{h} + \sum_{\substack{k=0 \\ k\neq i}}^{\infty} \frac{P_{ik}(h) \; P_{kj}(t)}{h}\right] \; . \qquad (2.45)$$

The left hand side is the derivative of $P_{ij}(t)$ with respect to time. With the assumption that the limit and summation can be interchanged one has,

$$\frac{dP_{ij}(t)}{dt} = \lim_{h\to 0} \left[\frac{P_{ii}(h)-1}{h}\right] P_{ij}(t) + \sum_{\substack{k=0 \\ k\neq i}}^{\infty} \lim_{h\to 0} \left[\frac{P_{ik}(h)}{h}\right] P_{kj}(t) \; . \qquad (2.46)$$

Now let

$$q_{ii} = \lim_{h\to 0} \left[\frac{P_{ii}(h)-1}{h}\right] , \qquad (2.47)$$

and

$$q_{ik} = \lim_{h \to 0} \left[\frac{P_{ik}(h)}{h} \right] \quad \text{for} \quad k \neq i. \tag{2.48}$$

Substituting (2.47) and (2.48) into (2.46) gives

$$\frac{dP_{ij}(t)}{dt} = q_{ii} \, P_{ij}(t) + \sum_{\substack{k=0 \\ k \neq i}}^{\infty} q_{ik} \, P_{kj}(t) \ . \tag{2.49}$$

This equation is known as the Chapman-Kolmogorov backwards differential equation.

Equations (2.47) and (2.48) have the following interpretations: For small values of h the probability that the process makes a transition out of state i is approximately $1+q_{ii}h$. More precisely,

$$P_{ii}(h) = 1 + q_{ii}h + o(h) \ . \tag{2.50}$$

The notation $o(h)$ represents any function that goes to zero faster than h. That is,

$$\lim_{h \to 0} \frac{o(h)}{h} = 0 \ . \tag{2.51}$$

It should be obvious that q_{ii} must be negative. In fact $q_{ii}=-\xi_i$. To show this, recall that the time spent in state i is exponentially distributed with mean $1/\xi_i$. Hence for small values of h

$$\begin{aligned} P_{ii}(h) &= P[T_i > h] = e^{-\xi_i h} \\ &= 1 - \xi_i h + \frac{(\xi_i h)^2}{2!} - \frac{(\xi_i h)^3}{3!} + \cdots \\ &= 1 - \xi_i h + o(h), \end{aligned} \tag{2.52}$$

and

$$o(h) = \frac{(\xi_i h)^2}{2!} - \frac{(\xi_i h)^3}{3!} + \cdots \quad . \tag{2.53}$$

The interpretation is that when the process is in state i, it departs at mean rate $-q_{ii}$.

Similarly for small values of h the probability that the process makes a transition to state k, given that it is currently in state i, can be approximated by q_{ik} h. More precisely,

$$P_{ik}(h) = q_{ik}\, h + o(h) \ . \tag{2.54}$$

The interpretation is that when the process is in state i, the rate of flow to state k is q_{ik}. For small values of h, $P_{ik}(h)$ is the probability that there is a transition from state i, and the transition is to state k. Since these two events are independent

$$\begin{aligned} P_{ik}(h) &= P[T_i \le h]\ p_{ik} = (1 - e^{-\xi_i h})\ p_{ik} \\ &= (1 - (1 - \xi_i h + \frac{(\xi_i h)^2}{2!} - \frac{(\xi_i h)^3}{3!} + \cdots))\ p_{ik} \\ &= (\xi_i h + o(h))\ p_{ik} \\ &= (\xi_i\ p_{ik}\ h) + o(h) \end{aligned} \tag{2.55}$$

where p_{ik} is the probability the transition is from i to k, and again o(h) picks up all terms with powers of h. Hence,

$$q_{ik} = \xi_i\ p_{ik} \ . \tag{2.56}$$

Since ξ_i is the conditional rate of flow from state i, then $\xi_i\ p_{ik}$ must be the conditional rate of flow from i to j. It follows that

$$q_{ii} = \sum_{\substack{k=0 \\ k \neq i}}^{\infty} q_{ik} \quad , \quad \text{and} \quad \sum_{k=0}^{\infty} q_{ik} = 0 \; . \tag{2.57}$$

As in the discrete case, there is also a forward Chapman-Kolmogorov equation. The forward differential equation is derived by interchanging t and h in Equation (2.44), writing out the k=j term and following the procedure for backwards differential equation. The results is

$$\frac{dP_{ij}(t)}{dt} = P_{ij}(t)\; q_{jj} \; + \sum_{\substack{k=0 \\ k \neq j}}^{\infty} P_{ik}(t)\; q_{kj} \tag{2.58}$$

where

$$q_{kj} = \lim_{h \to 0} \left[\frac{P_{kj}(h)}{h}\right] \;\; \text{for } j \neq k, \;\; \text{and} \;\; q_{jj} = \lim_{h \to 0} \left[\frac{P_{jj}(h)-1}{h}\right] .$$

Both the backwards *(2.49)* and *forward (2.58)* equations define a set of differential equations which can be put into vector form by defining [Q] as

$$[Q] = \lim_{h \to 0} \left[\frac{[P(t)]-[I]}{h}\right] \; , \tag{2.59}$$

where [I] is the identity matrix. The matrix [Q] is called the rate matrix. The resulting backward matrix equation is

$$\frac{d[P(t)]}{dt} = [P(t)]\; [Q] \; , \tag{2.60}$$

and the resulting forward matrix equation is

$$\frac{d[P(t)]}{dt} = [Q]\; [P(t)] \; . \tag{2.61}$$

Since both equations describe the same process, they both have the same solutions. The initial conditions for the equations are

$$P_{ij}(0) = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ if } i\neq j \end{cases} \qquad (2.62)$$

or in matrix form $[P(0)] = [I]$. The initial conditions simply state that if the process is in state i at any given time, it is also in state i zero time units later.

Hence, a set of differential equations can be written and the transition probabilities calculated. The solution to the matrix equation is

$$[P(t)] = e^{[Q]t}$$

where

$$e^{[Q]t} = [I] + [Q]t + \frac{[Q]^2t^2}{2!} + \frac{[Q]^3t^3}{3!} + \cdots \quad . \qquad (2.63)$$

These equations provide a formal solution. In practice this method is often so cumbersome that alternate methods must be sought [CLAR70].

Recall that the primary purpose of finding $P_{ij}(t)$ was to determine $P_j(t)$. That is,

$$P_j(t) = \sum_{i=0}^{\infty} P_{ij}(t)\ P_i(0) \ . \qquad (2.64)$$

The derivative of this equation is

$$\frac{dP_j(t)}{dt} = \sum_{i=0}^{\infty} \left[\frac{dP_{ij}(t)}{dt}\right] P_i(0) \ . \qquad (2.65)$$

Substituting Equation (2.58) for $\frac{dP_{ij}(t)}{dt}$, yields

$$\frac{dP_j(t)}{dt} = \sum_{i=0}^{\infty} \left[P_{ij}(t)\ q_{jj} + \sum_{\substack{k=0 \\ k \neq j}}^{\infty} P_{ik}(t)\ q_{kj} \right] P_i(0)$$

$$= \left[\sum_{i=0}^{\infty} P_{ij}(t)\ P_i(0) \right] q_{jj} + \sum_{\substack{k=0 \\ k \neq j}}^{\infty} \left[\sum_{i=0}^{\infty} P_{ik}(t)\ P_i(0) \right] q_{kj} \ . \qquad (2.66)$$

From Equation (2.42) the first term in brackets is $P_j(t)$ and the second term is $P_k(t)$. Therefore,

$$\frac{dP_j(t)}{dt} = P_j(t)\ q_{jj} + \sum_{\substack{k=0 \\ k \neq j}}^{\infty} P_k(t)\ q_{kj} \ , \qquad (2.67)$$

or in vector form

$$\frac{dV(t)}{dt} = V(t)\ [Q] \ . \qquad (2.68)$$

2.5.1 The Poisson Process

Consider the problem of determining the number of arrivals that occur in an interval of time t, given that the interarrival times (time between arrivals) are exponentially distributed with a mean of $1/\xi$. Since only arrivals can occur $P_{ij}(t) = 0$ when $j<i$ (departures are not allowed). Since the probability of multiple arrivals in an infinitesimal interval h is o(h), it follows that the rates are :

$$q_{kj} = \begin{cases} 0 \text{ when } k \neq j-1 \\ \xi \text{ when } k = j-1 \end{cases} \qquad (2.69)$$

and

$$q_{jj} = -\xi \ . \qquad (2.70)$$

The forward differential equations are :

$$\frac{dP_{ii}(t)}{dt} = -\xi \; P_{ii}(t) \qquad j=i \qquad (2.71)$$

$$\frac{dP_{ij}(t)}{dt} = -\xi \; P_{ij}(t) + \xi \; P_{i,j-1}(t) \qquad j\geq i+1 \; . \qquad (2.72)$$

The solution to the j=i equation is obviously

$$P_{ii}(t) = e^{-\xi t}. \qquad (2.73)$$

The results of the j=1 equation can be substituted into the j=i+1 equation to obtain

$$\frac{P_{i,i+1}(t)}{dt} = -\xi \; P_{i,i+1}(t) + \xi e^{-\xi t} \; , \qquad (2.74)$$

or

$$\frac{P_{i,i+1}(t)}{dt} + \xi \; P_{i,i+1}(t) = \xi e^{-\xi t} \; . \qquad (2.75)$$

The solution is easily obtained by taking the Laplace transforms of both sides. More precisely,

$$P_{i,i+1}(S) = \frac{\xi}{(S+\xi)^2} \qquad (2.76)$$

and

$$P_{i,i+1}(t) = \xi t e^{-\xi t} \; . \qquad (2.77)$$

It follows by induction that

$$P_{ij}(S) = \frac{\xi^{j-i}}{(S+\xi)^{j-i+1}} \; , \qquad (2.78)$$

and

$$P_{ij}(t) = \frac{(\xi t)^{j-i} \; e^{\xi t}}{(j-i)!} \; . \qquad (2.79)$$

This last equation is the celebrated Poisson process. Thus, the Poisson process is a special case of the Markov process. It is important to emphasize that this also implies that for a Poisson process the time between arrivals is exponentially distributed. The reverse is also true, if the time between arrivals is exponentially distributed the process is Poisson.

2.5.2 Continuous Time - Limiting Probabilities

It was shown in the case of the discrete-time Markov process that under certain conditions the limiting probabilities existed, and were independent of the initial state, that is,

$$P_j = \lim_{n\to\infty} P_{ij}{}^n = \lim_{n\to\infty} P_j{}^n \ . \qquad (2.14)$$

Recalling that the difference between $P_{ij}{}^n$ and $P_{ij}(t)$ is that the discrete parameter n is replaced by the continuous parameter t, it seems plausible that for the continuous-time process

$$P_j = \lim_{t\to\infty} P_{ij}(t) = \lim_{t\to\infty} P_j(t) \ . \qquad (2.80)$$

This is indeed the case, and the conditions for the limit to exist are the same as those for the discrete case.

To derive a set of equations for P_j, it is necessary to take the limit as t approaches infinity of both the backward and forward Chapman-Kolmogorov differential equation. The backward equation results in

$$\lim_{t\to\infty} \frac{dP_{ij}(t)}{dt} = q_{ii} \lim_{t\to\infty} P_{ij}(t) + \lim_{t\to\infty} \sum_{\substack{k=0\\k\neq i}}^{\infty} q_{ik} P_{kj}(t) \quad . \qquad (2.81)$$

If the limit and summation can be interchanged then

$$\lim_{t\to\infty} \frac{dP_{ij}(t)}{dt} = q_{ii}P_j + \sum_{\substack{k=0\\ K\neq i}}^{\infty} q_{ik}\, P_j \ ,$$

$$= P_j \sum_{k=0}^{\infty} q_{ik}$$

$$= 0 \ . \tag{2.82}$$

By applying the same procedure to the forward equation and using the results of the backward equation, one obtains

$$\lim_{t\to\infty} \frac{dP_{ij}(t)}{dt} = \lim_{t\to\infty} P_{ij}\ q_{jj} + \lim_{t\to\infty} \sum_{\substack{k=0\\ k\neq j}}^{\infty} P_{ik}(t)\ q_{kj} = 0 \tag{2.83}$$

$$P_j\ q_{jj} + \sum_{\substack{k=0\\ k\neq j}}^{\infty} P_k\ q_{kj} = 0 \tag{2.84}$$

$$\sum_{k=0}^{\infty} P_k\ q_{kj} = 0 \tag{2.85}$$

The vector form of Equation (2.85) is

$$V\ [Q] = 0 \ . \tag{2.86}$$

Equation (2.84), or (2.85), or (2.86) along with the conservation of probability equation,

$$\sum_{k} P_k = 1 \ , \tag{2.87}$$

uniquely determine the limiting probabilities.

Note that the same results could have been obtained from Equation

(2.67) by taking the limit as n approaches infinity and setting

$$\lim_{t \to \infty} \frac{dP_j(t)}{dt} = 0 \ . \qquad (2.88)$$

The interpretation of Equation (2.84) is important. Recall that $(-q_{jj})$ is the rate of flow from state j when the the process is in state j. Since P_j is the proportion of time the process is in state j, it follows that

$$P_j\ (-q_{jj}) \ = \text{rate at which the process leaves state j.} \qquad (2.89)$$

Similarly, when the process is in state k it goes to state j at rate q_{kj}, therefore

$$\sum_{k \neq j} P_k\ q_{kj} = \ \text{rate the process enters state j} \ . \qquad (2.90)$$

Hence, Equation (2.84) is a statement of the equality of the rates at which the process enters and leaves state j. Because of this the limiting or steady-state probabilities are also referred to as the equilibrium probabilities.

CHAPTER 3

ELEMENTARY QUEUEING SYSTEMS IN EQUILIBRIUM

3.1 Introduction

An elementary queueing system is one in which both the interarrival and service times are exponentially distributed. These include a number of complex systems involving finite storage, multiple servers, finite customer populations, and the like. All of these fall into the category of birth and death processes.

3.2 Birth and Death Processes

A birth and death process is a continuous-time Markov process such that: (1) the state represents the number of persons in the system,k, (2) new arrivals enter at an exponential rate λ_k, and (3) people depart the system at an exponential rate μ_k. That is whenever there are k persons in the system, the time until the next arrival is exponentially distributed with mean $1/\lambda_k$ and is independent of the time until the next departure which is itself exponentially distributed with mean $1/\mu_k$. Thus, a birth and death process is a continuous-time Markov process with states {0,1,2,..} for which transitions from state k may go only to either state k+1 or state k-1.

In terms of the rates in the last chapter :

$$q_{k,k+1} = \lambda_k \tag{3.1}$$

$$q_{k,k-1} = \mu_k. \tag{3.2}$$

The nearest-neighbor condition requires that $q_{kj}=0$ for $|k-j| > 1$.

Moreover, since

$$\sum_j q_{kj} = 0, \text{ then } q_{kk} = -(\lambda_k+\mu_k). \qquad (3.3)$$

The problem is to determine the limiting or steady-state probabilities. The equations derived in the last chapter could be used, however, the derivation is straightforward and follows from first principles. Therefore, rather than use the results of the last chapter which tend to camouflage the basic approach, the equations will be rederived for this simpler case.

The probability that the system contains k persons at time t+h can be expressed as

$$P_k(t+h) = \sum_{i=0}^{\infty} P_i(t)\ P_{ik}(h) \ . \qquad (3.4)$$

If it is assumed that the probability of two or more state changes in infinitesimal time h is negligible compared to single state change, then Equation(3.4) becomes

$$P_k(t+h) = P_k(t)\ P_{kk}(h) + P_{k-1}(t)\ P_{k-1,k}(h)$$
$$+ P_{k+1}(t)\ P_{k+1,k}(h) + o(h) \qquad (3.5)$$

where

$P_{kk}(h)$ = P[zero arrivals and zero departures in h| k in population],

$P_{k-1,k}(h)$ = P[one arrival and zero departures in h| k-1 in population],

$P_{k+1,k}(h)$ = P[zero arrivals and one departure in h| k+1 in population],

o(h) = P[multiple arrivals and/or multiple departures in h].

In order to justify the o(h) assumption and to calculate these probabilities, it is first necessary to find the individual arrival and departure probabilities. The following assumes that the process is in state k. The birth probabilities are:

$$
\begin{aligned}
P[\text{zero births in } h] &= P[T_b > h] = 1 - P[T_b \le h] \\
&= 1 - (1 - e^{-\lambda_k h}) \\
&= e^{-\lambda_k h} \\
&= 1 - \lambda_k h + \frac{(\lambda_k h)^2}{2!} - \frac{(\lambda_k h)^3}{3!} + \cdots \\
&= 1 - \lambda_k h + o(h). \qquad (3.6)
\end{aligned}
$$

$$
\begin{aligned}
P[\text{one birth in } h] &= P[T_b \le h] \\
&= 1 - e^{-\lambda_k h} \\
&= 1 - \left[1 - \lambda_k h + \frac{(\lambda_k h)^2}{2!} - \frac{(\lambda_k h)^3}{3!} + \cdots \right] \\
&= \lambda_k h + o(h). \qquad (3.7)
\end{aligned}
$$

$$
\begin{aligned}
&P[\text{two or more births in } h] \\
&\quad = 1 - P[\text{zero arrivals in } h] - P[\text{one arrival in } h] \\
&\quad = 1 - [1 - \lambda_k h + o(h)] - [\lambda_k h + o(h)] \\
&\quad = o(h). \qquad (3.8)
\end{aligned}
$$

The death calculations are the same as births except the parameter λ_k is replaced by μ_k. More precisely,

$$P[\text{zero deaths in } h] = 1 - \mu_k h + o(h) \qquad (3.9)$$

$$P[\text{one death in } h] = \mu_k h + o(h) \qquad (3.10)$$

$$P[\text{two or more deaths in } h] = o(h) \qquad (3.11)$$

The desired joint probabilities can now be calculated:

$$\begin{aligned}P[\text{zero births and zero deaths in } h] &= [1-\lambda_k h+o(h)]\ [1-\mu_k h+o(h)] \\ &= 1-\lambda_k h-\mu_k h+o(h). \qquad (3.12)\end{aligned}$$

$$\begin{aligned}P[\text{one birth and zero deaths in } h] &= [\lambda_k h+o(h)]\ [1-\mu_k h+o(h)] \\ &= \lambda_k h+o(h). \qquad (3.13)\end{aligned}$$

$$\begin{aligned}P[\text{zero births and one death in } h] &= [1-\lambda_k h+o(h)]\ [\mu_k h+o(h)] \\ &= \mu_k h+o(h). \qquad (3.14)\end{aligned}$$

$$P[\text{two or more arrivals and/or two or more departures}] = o(h). \qquad (3.15)$$

Substituting these results into Equatio. (3.5) and adjusting the subscripts to account for states k+1 and k-1 result in:

$$\begin{aligned}P_k(t+h) = [1-\lambda_k h-\mu_k h+o(h)]\ P_k(t) + [\lambda_{k-1} h+o(h)]\ P_{k-1}(t) \\ + [\mu_{k+1} h+o(h)]\ P_{k+1}(t) + o(h)\ . \qquad (3.16)\end{aligned}$$

Following the usual procedure of subtracting $P_k(t)$ from both sides, dividing by h, and taking the limit as h approaches zero, one obtains

$$\frac{dP_k(t)}{dt} = -(\lambda_k+\mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) \qquad (3.17)$$

Although Equation (3.17) is valid for all values of k, it is sometimes more convenient to separate it into two equations, one for k=0 and one for $k \geq 1$. This is because some of the terms are zero when k=0. That is, it is impossible to have a negative number of customers, and when the number of customers is zero the death rate is zero. More precisely,

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \qquad \text{for } k=0 \qquad (3.18a)$$

$$\frac{dP_k(t)}{dt} = -(\lambda_k+\mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) \qquad \text{for } k\geq 1 \qquad (3.18b)$$

Equations (3.18a) and (3.18b) are identical to the equations that would have been obtained by substituting the proper values of q_{kj} into Equation (2.67).

The solution to this set of differential equations depends on the rates λ_k and μ_k. Unfortunately, no matter how simple λ_k and μ_k are, it is nearly impossible to obtain the transient solution. Fortunately, one is usually only interested in the steady-state solution, and it is easy to obtain. The limits as t approaches infinity of Equations (3.18a) and (3.18b) are

$$0 = \lambda_0 P_o + \mu_1 P_1 \qquad \text{for } k=0 \qquad (3.19a)$$

$$0 = -(\lambda_k+\mu_k)P_k + \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} \qquad \text{for } k\geq 1, \qquad (3.19b)$$

or

$$\lambda_0 P_0 = \mu_1 P_1 \qquad \text{for } k=0 \qquad (3.20a)$$

$$(\lambda_k+\mu_k)P_k = \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{K+1} \qquad \text{for } k\geq 1\ . \qquad (3.20b)$$

The left hand sides of equations (3.20a) and (3.20b) are simply the rate of flow out of state k, while the right hand sides are the rate of flow into state k. In problems of this sort it is often helpful to sketch a diagram showing the average rate of flow from one state to another. Such a state-transition-rate diagram is shown in Figure 3.1.

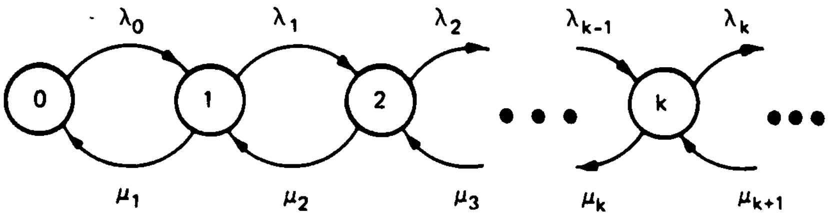


Figure 3.1 State-Transition-Rate Diagram of the Birth and Death Process.

By equating the flow rate out of a state to the flow rate into the same state, the following equations arise:

state	rate out = rate in	
0	$\lambda_0 P_0 = \mu_1 P_1$	
1	$(\lambda_1+\mu_1)P_1 = \lambda_0 P_0+\mu_2 P_2$	
2	$(\lambda_2+\mu_2)P_2 = \lambda_1 P_1+\mu_3 P_3$	
3	$(\lambda_3+\mu_3)P_3 = \lambda_2 P_2+\mu_4 P_4$	
k,k≥1	$(\lambda_k+\mu_k)P_k = \lambda_{k-1}P_{k-1}+\mu_{k+1}P_{k+1}$.	(3.21)

Solving these equations in terms of P_0 yields:

$$P_1 = \frac{\lambda_0}{\mu_1} P_0$$

$$P_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2} P_0$$

$$P_3 = \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} P_0$$

$$P_k = \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{k-2}\lambda_{k-1}}{\mu_1\mu_2\mu_3\cdots\mu_{k-1}\mu_k} P_0. \qquad (3.22)$$

P_0 can be determined from the conservation of probabilities equation. That is

$$\sum_{k=0}^{\infty} P_k = 1$$

$$P_0 + P_0 \sum_{k=1}^{\infty} \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{k-2}\lambda_{k-1}}{\mu_1\mu_2\mu_3\cdots\mu_{k-1}\mu_k} = 1$$

$$P_0 = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{k-2}\lambda_{k-1}}{\mu_1\mu_2\mu_3\cdots\mu_{k-1}\mu_k}\right]^{-1}. \qquad (3.23)$$

Clearly for the limiting probabilities to exist it is necessary that

$$\sum_{k=1}^{\infty} \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{k-2}\lambda_{k-1}}{\mu_1\mu_2\mu_3\cdots\mu_{k-1}\mu_k} < \infty . \qquad (3.24)$$

This condition also may be shown to be sufficient [KLEI75].

3.3 Birth and Death Processes and Elementary Queueing Systems

By properly selecting the birth and death rates, a number of rather complex queueing systems can be modeled. The birth process corresponds to the arrival process and death process to the service process. If the system is in state k, the arrival rate is λ_k and the service rate μ_k. The birth and death model does not explicitly allow for specifying a rule for deciding which among several customers is to be served next, nor is one required. This is because the state of the process is invariant to the order in which customers are served. The reason for this is that all customers are assumed to be statistically identical and the service process has the memoryless property. Perhaps the most commonly used rule and the one whose operation is most easily visualized, is the first-come-first-serve rule. However the service discipline can be last-come-first-serve, or service in random order. An arriving customer can even preempt the service of a customer without changing the dynamics. The key in all these cases is that a server is never idle when customers are waiting in the queue, and that the probability of a customer departing the service center in the next incremental interval is $\mu_k h$ independent of the amount of service previously received. Of course, it is necessary that the service discipline be work conservative. However, this does not imply that if a customer is interrupted the amount of service he received must be remembered. The PDF of the service time is memoryless!

3.4 M/M/1

The simplest and one of the most celebrated queueing systems is one in which the birth and death rates are constant regardless of the state of the system. More precisely,

$$\lambda_k = \lambda \quad \text{for } k = 0,1,2,\ldots \tag{3.25}$$

$$\mu_k = \mu \quad \text{for } k = 1,2,3,\ldots\ . \tag{3.26}$$

Recall that when the arrival rate is constant (that is it does not depend upon the state of the process) the arrival process is Poisson.

The state-transition-rate diagram of this system is depicted in Figure 3.2.

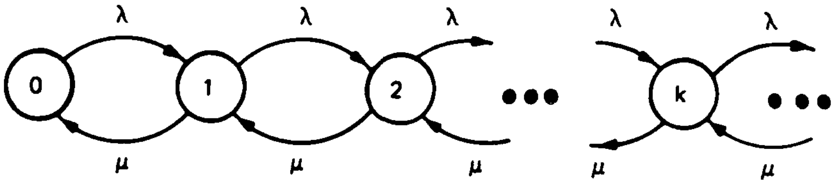


Figure 3.2 State-Transition-Rate Diagram for M/M/1 System.

By equating the flow rate out of a state to the flow rate into the same state, the following equations arise :

state	rate out = rate in	
0	$\lambda P_0 = \mu P_1$	
1	$(\lambda+\mu)P_1 = \lambda P_0+\mu P_2$	
2	$(\lambda+\mu)P_2 = \lambda P_1+\mu P_3$	
3	$(\lambda+\mu)P_3 = \lambda P_2+\mu P_4$	
$k, k\geq 1$	$(\lambda+\mu)P_k = \lambda P_{k-1}+\mu P_{k+1}$.	(3.27)

Solving these equations in terms of P_0 yields :

$$P_1 = (\lambda/\mu)\ P_0$$

$$P_2 = (\lambda/\mu)^2\ P_0$$

$$P_3 = (\lambda/\mu)^3\ P_0$$

$$P_k = (\lambda/\mu)^k\ P_0. \qquad (3.28)$$

Again, P_0 can be found from the conservation of probabilities equation. More precisely,

$$\sum_{k=0}^{\infty} P_k = 1$$

and

$$P_0 + \sum_{k=1}^{\infty} (\lambda/\mu)^k \, P_0 = 1$$

$$P_0 \sum_{k=0}^{\infty} (\lambda/\mu)^k = 1$$

$$P_0 \; \frac{1}{1-(\lambda/\mu)} = 1$$

$$P_0 = 1-(\lambda/\mu). \qquad (3.29)$$

The utilization of the service center is

$$\rho = 1-P_0 = \lambda/\mu. \qquad (3.30)$$

The quantity λ/μ appears frequently in the performance parameters, and therefore it is customary to give them in terms of the utilization, ρ.

The steady-state probability that the system contains k customers is

$$P_k = \rho^k(1-\rho). \qquad (3.31)$$

The mean number of customers in the system and its variance can be calculated by the probability generating function (which is very similar to the z-transform) [KLEI75] [KOBA81]. By definition the probability generating function is,

$$P(z) = \sum_{k=0}^{\infty} P_k \, z^k \; . \qquad (3.32)$$

The mean or first moment is is equal to derivative of P(z) evaluated at z equal to one. More precisely,

$$\frac{d}{dz} \Big[\sum_{k=0}^{\infty} P_k \, z^k \Big]_{z=1} = \sum_{k=0}^{\infty} \frac{d}{dz} \Big[P_k \, z^k \Big]_{z=1} = \sum_{k=0}^{\infty} k \; P_k = E[N] = L \quad . \qquad (3.33)$$

The second derivative of P(z) evaluated at z equal to one is

$$\sum_{k=0}^{\infty} k\ (k-1)\ P_k \;=\; E[N^2] - E[N]\ . \qquad (3.34)$$

Hence, the second moment and variance are respectfully:

$$E[N^2] = \left.\frac{d^2P(z)}{dz^2}\right|_{z=1} + E[N]\ , \qquad (3.35)$$

$$\mathrm{Var}[N] = E[N^2]-(E[N])^2. \qquad (3.36)$$

Returning to the problem at hand and substituting $P_k=\rho^k(1-\rho)$ results in

$$P(z) = \sum_{k=0}^{\infty} \rho^k(1-\rho)\ z^k = (1-\rho)\sum_{k=0}^{\infty}(\rho z)^k$$

$$= \frac{1-\rho}{1-\rho z}\ , \qquad (3.37)$$

and

$$L = \frac{\rho}{1-\rho}\ , \qquad (3.38)$$

and

$$\mathrm{Var}[N] = \frac{\rho}{(1-\rho)^2}\ . \qquad (3.39)$$

The average response time can be calculated by Little's law. More precisely,

$$R = \frac{L}{T} = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu(1-\rho)}\ . \qquad (3.40)$$

Little's law can also be used to calculate the mean number of customers in the queue, L_q. However, it is first necessary to calculate the mean waiting time, W_q. Since the mean of a sum is equal to the sum of the means (irrespective of dependencies involved), it follows that

$$R = W_q + E[S], \tag{3.41}$$

and

$$W_q = R - E[S] = \frac{1}{\mu(1-\rho)} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} \quad . \tag{3.42}$$

Little's law can now be applied to determine L_q:

$$L_q = \lambda\, W_q = \frac{\lambda}{\mu}\;\frac{\rho}{(1-\rho)} \;=\; \frac{\rho^2}{1-\rho} \;. \tag{3.43}$$

Another interesting quantity to calculate is the probability of finding at least i customers in the system:

$$P[k\geq i] = \sum_{k=i}^{\infty} P_k \;=\; \sum_{k=i}^{\infty} \rho^k(1-\rho) = (1-\rho)\left[\sum_{k=0}^{\infty}\rho^k - \sum_{k=0}^{i-1}\rho^k\right] = \rho^i \tag{3.44}$$

where the last expression follows upon application of the algebraic identity

$$\sum_{k=0}^{i-1} \rho^k = (1-\rho^i)/(1-\rho). \tag{3.45}$$

Figure 3.3 compares the normalized mean response time ($1/\mu=1$) of the M/M/1 system to that of the D/D/1 system. For both systems the end-point values are the same. More precisely, when $\rho=0$, the normalized mean response time is one, and when $\rho=1$, the response time is infinite. However, for values of ρ near one there is an extreme difference. For

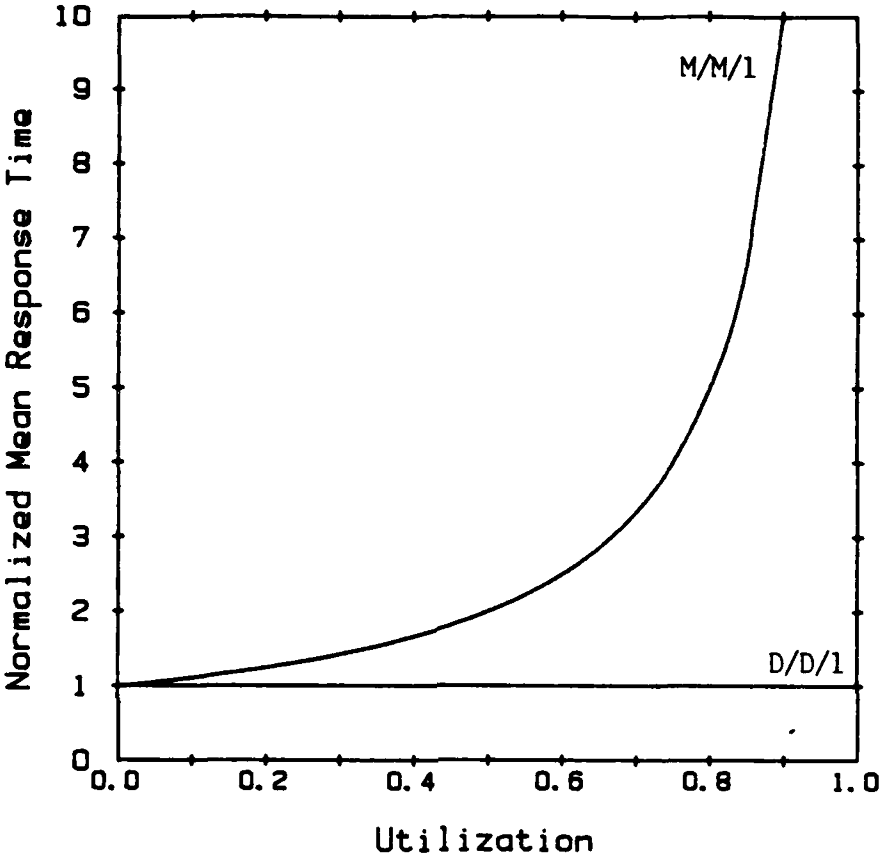


Figure 3.3 Normalized Mean Response Time Versus Utilization for M/M/1 and D/D/1 Systems.

example, when ρ=0.9 the mean response time of the M/M/1 system is ten times that of the D/D/1 system. Clearly, the D/D/1 curve is the optimal one, and a large penalty is paid for operating the M/M/1 system near its maximum capacity. The reason for this is that there are no statistical fluctuations in the D/D/1 system, whereas both the interarrival and service times are random variables in the M/M/1 system. Any reduction in the variation of either of these reduces the response time, while any increase results in an increased response time. In fact it will be shown later that the mean waiting time (a quantity closely related to the response time) of the M/D/1 system is exactly one-half of that for the M/M/1 system.

Figure 3.4 compares the mean number of customers in the M/M/1 system to that of the D/D/1 system. Note that curve for the D/D/1 is a straight line over the region $\rho<1$ ($L=\rho$), and when $\rho=1$ the number of customers is infinite. Again the differences in the curves are due to the statistical fluctuations in M/M/1 system.

3.5 M/M/m - Finite Number of Servers

Now consider the case when the number of servers is more than one and finite. Assume there are m servers. The birth and deaths rates are:

$$\lambda_k = \lambda \quad k = 0,1,2,... \tag{3.46}$$

$$\mu_k = \min[k\mu, m\mu] \ . \tag{3.47}$$

The state-transition-rate diagram is depicted in Figure 3.5.

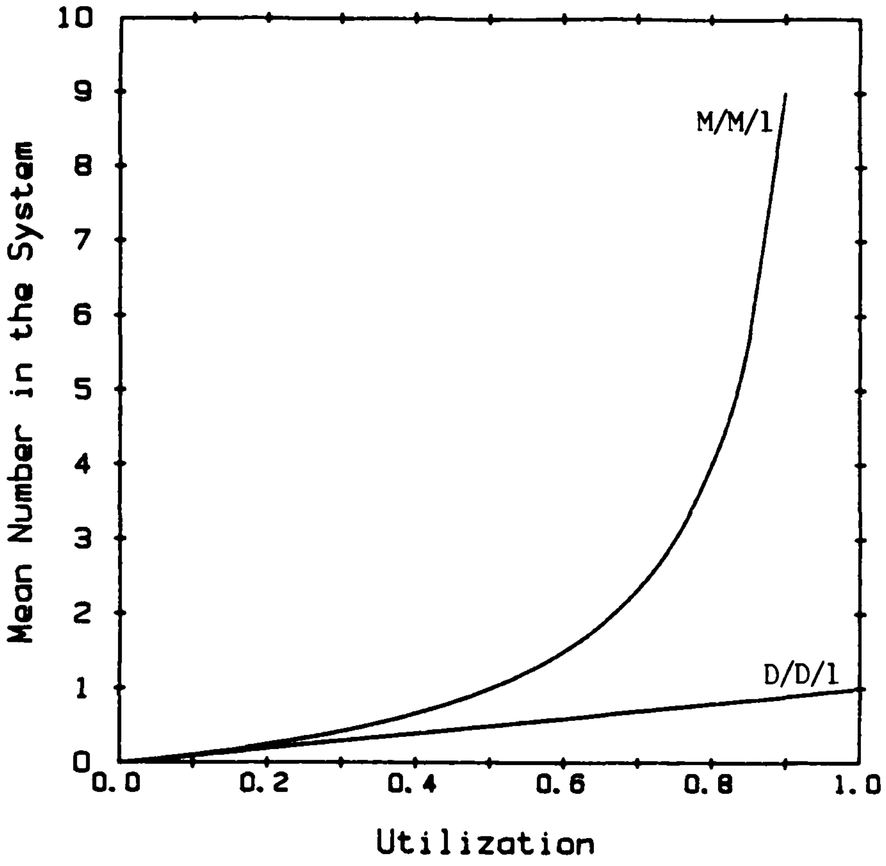


Figure 3.4 Mean Number of Customers Versus Utilization for the M/M/1 and D/D/1 Systems.

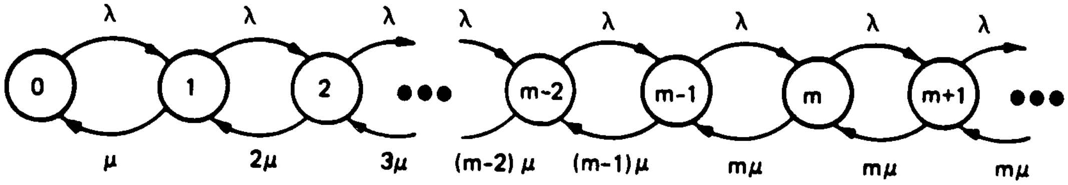


Figure 3.5. State-Transition-Rate Diagram for M/M/m System.

The steady-state equations are :

state	rate out = rate in
0	$\lambda P_0 = \mu P_1$
1	$(\lambda+\mu)P_1 = \lambda P_0 + 2\mu P_2$
2	$(\lambda+2\mu)P_2 = \lambda P_1 + 3\mu P_3$
3	$(\lambda+3\mu)P_3 = \lambda P_2 + 4\mu P_4$
m-1	$[\lambda+(m-1)\mu]P_{m-1} = \lambda P_{m-2} + m\mu P_m$
m	$(\lambda+m\mu)P_m = \lambda P_{m-1} + m\mu P_{m+1}$
m+1	$(\lambda+m\mu)P_{m+1} = \lambda P_m + m\mu P_{m+2}$
m+2	$(\lambda+m\mu)P_{m+2} = \lambda P_{m+1} + m\mu P_{m+3}$. (3.48)

Solving the first m equations in terms of P_0 yields:

$$P_1 = (\lambda/\mu)\ P_0$$

$$P_2 = (1/2)\ (\lambda/\mu)^2\ P_0$$

$$P_3 = (1/6)\ (\lambda/\mu)^3\ P_0\ ,$$

and

$$P_k = (1/k!)\ (\lambda/\mu)^k\ P_0 \qquad \text{for } k\leq m\ . \qquad (3.49)$$

Similarly, solving the equations for states m,m+1, and m+2 yields

$$P_{m+1} = (1/m!m)\ (\lambda/\mu)^{m+1}\ P_0$$

$$P_{m+2} = (1/m!m^2)\ (\lambda/\mu)^{m+2}\ P_0$$

$$P_{m+3} = (1/m!m^3)\ (\lambda/\mu)^{m+3}\ P_0\ ,$$

and

$$P_k = (1/m!m^{k-m})\ (\lambda/\mu)^k\ P_0 \qquad \text{for } k \geq m\ . \qquad (3.50)$$

Collecting the results together

$$P_k = \begin{cases} (1/m!m^{k-m})\ (\lambda/\mu)^k\ P_0 & \text{for } k \geq m \\ (1/k!)\ (\lambda/\mu)^k\ P_0 & \text{for } k \leq m\ , \end{cases} \qquad (3.51)$$

or equivalently

$$P_k = \begin{cases} \dfrac{(m\rho)^k}{k!}\ P_0 & \text{for } k \leq m\ , \\ \dfrac{m^m\ \rho^k}{m!}\ P_0 & \text{for } k \geq m\ , \end{cases} \qquad (3.52)$$

where $\rho = \dfrac{\lambda}{m\mu}$.

Solving for P_0 in the usual way results in:

$$P_0 \left[\ 1 + \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \sum_{k=m}^{\infty} \frac{m^m\ \rho^k}{m!}\ \right] = 1$$

$$P_0 \left[\ \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \sum_{k=0}^{\infty} \rho^k\ \right] = 1$$

$$P_0 = \left[\ \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!\ (1-\rho)}\ \right]^{-1} . \qquad (3.53)$$

Similarly the mean number of customers in the system is:

$$L = \sum_{k=1}^{\infty} k\, P_k$$

$$= \sum_{k=1}^{m-1} \frac{k\ (m\rho)^k}{k!} P_0 + \sum_{k=m}^{\infty} \frac{k\ m^m\ \rho^k}{m!} P_0$$

$$= P_0 \sum_{k=1}^{m-1} \frac{(m\rho)^k}{(k-1)!} + P_0 \frac{(m\rho)^m}{m!} \sum_{k=0}^{\infty} (m+k)\ \rho^k$$

$$= P_0\ (m\rho) \sum_{k=0}^{m-2} \frac{(m\rho)^k}{k!} + P_0 \frac{(m\rho)^m}{m!} \left[\frac{m}{(1-\rho)} + \frac{\rho}{(1-\rho)^2} \right]$$

$$= m\rho + \frac{\rho\ (m\rho)^m}{m!\ (1-\rho)^2} P_0 \quad . \tag{3.54}$$

Figure 3.6 is a plot of the mean number of customers versus the arrival rate for m=1, m=2, and m=3 ($1/\mu=1$). Observe that the shape of all three curves are similar.

Again Little's Law can be used to determine the mean response time:

$$R = \frac{L}{T} = \frac{L}{\lambda} = \frac{1}{\mu} \left[1 + \frac{\rho\ (m\rho)^{m-1}}{m!\ (1-\rho)^2} P_0 \right] \quad . \tag{3.55}$$

Figure 3.7 is a plot of the normalized mean response time verse the arrival rate for m=1, m=2, and m=3. Again observe that the shape of the curves are very similar.

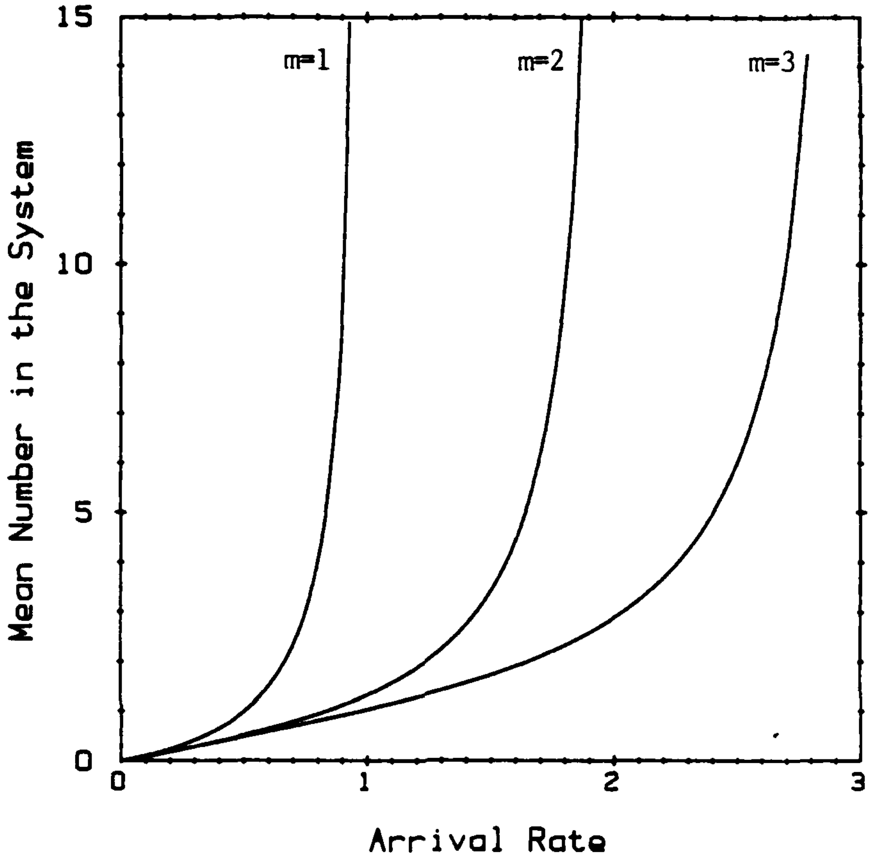


Figure 3.6 Mean Number of Customers Versus Arrival Rate for the M/M/m System.

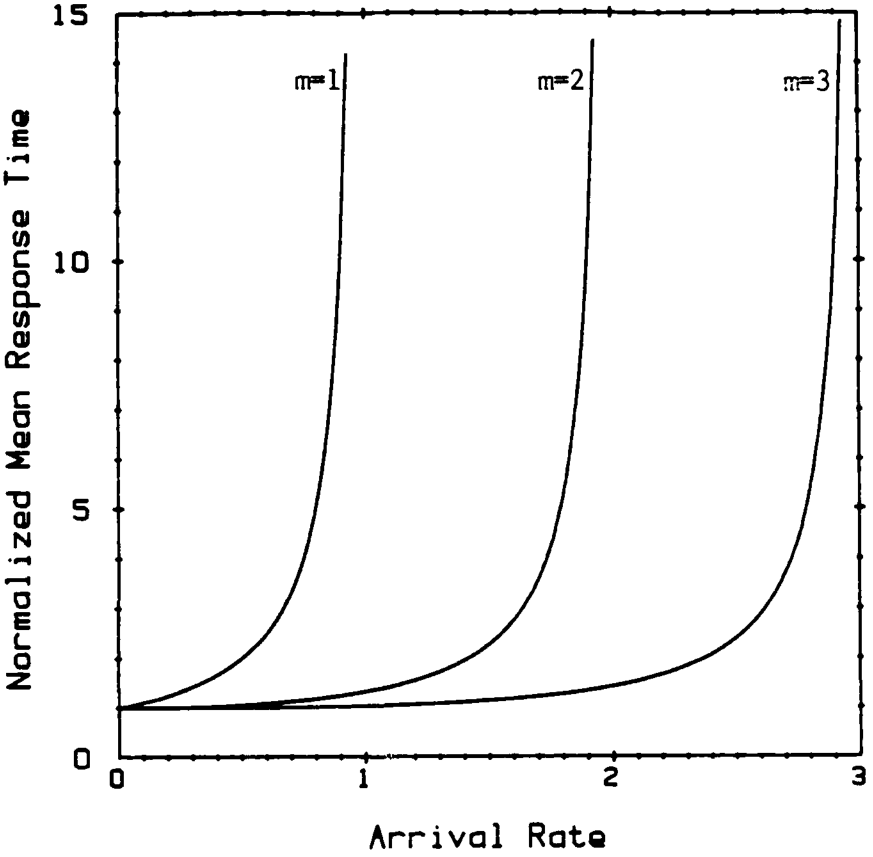


Figure 3.7 Mean Response Time Versus Arrival Rate for the M/M/m System.

3.6 M/M/∞ - Infinite Number of Servers

Now consider the case when the number of servers is infinite. That is, whenever a customer enters the queue, he immediately starts to receive service. This system is equivalent to one in which the number of servers always equals the number of customers in the system. In terms of the birth and death model the rates are :

$$\lambda_k = \lambda \qquad k = 0,1,2,... \tag{3.56}$$

$$\mu_k = k\mu \qquad k = 1,2,3,.... \; . \tag{3.57}$$

The state-transition-rate diagram is depicted in Figure 3.8.

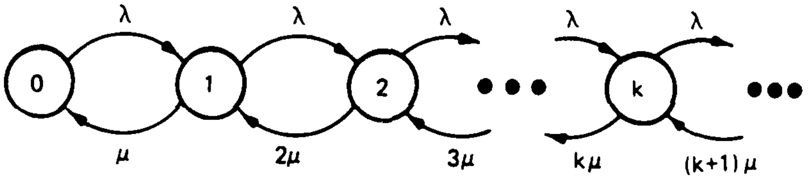


Figure 3.8 State-Transition-Rate Diagram for M/M/∞ System.

The resulting equilibrium equations are:

state	rate out = rate in
0	$\lambda P_0 = \mu P_1$
1	$(\lambda+\mu)P_1 = \lambda P_0+2\mu P_2$
2	$(\lambda+2\mu)P_2 = \lambda P_1+3\mu P_3$
3	$(\lambda+3\mu)P_3 = \lambda P_2+4\mu P_4$
$k, k\geq 1$	$(\lambda+k\mu)P_k = \lambda P_{k-1}+(k+1)\mu P_{k+1}$. (3.58)

Solving these equation in terms of P_0 yields:

$$P_1 = (\lambda/\mu)\ P_0$$

$$P_2 = (1/2)\ (\lambda/\mu)^2\ P_0$$

$$P_3 = (1/6)\ (\lambda/\mu)^3\ P_0$$

$$P_k = (1/k!)\ (\lambda/\mu)^k\ P_0\ . \tag{3.59}$$

Solving for P_0 yields

$$P_0 \left[1 + \sum_{k=1}^{\infty} (1/k!)\ (\lambda/\mu)^k\right] = 1$$

$$P_0 \left[\sum_{k=0}^{\infty} (1/k!)\ (\lambda/\mu)^k\right] = 1$$

$$P_0\ e^{\lambda/\mu} = 1$$

$$P_0 = e^{-\lambda/\mu}. \tag{3.60}$$

Hence,

$$P_k = (1/k!)\ (\lambda/\mu)^k\ e^{-\lambda/\mu}. \tag{3.61}$$

The mean number of customers in the service center is:

$$L = \sum_{k=0}^{\infty} k\ P_k$$

$$= \sum_{k=0}^{\infty} k\ (1/k!)\ (\lambda/\mu)^k\ e^{-\lambda/\mu}$$

$$= e^{-\lambda/\mu}\ (\lambda/\mu) \sum_{k=0}^{\infty} (1/k!)\ (\lambda/\mu)^k$$

$$= \lambda/\mu\ . \tag{3.62}$$

The throughput of the system is obviously λ. The mean response time can easily be determined by Little's law:

$$R = \frac{L}{\lambda} = \frac{1}{\mu} \, . \tag{3.63}$$

This is obviously correct since each arriving customer is immediately granted a server and the average service time is $1/\mu$.

3.7 M/M/1/K - Finite Storage

Now consider the problem in which the arrival rate and departure rates are constant, but there is a maximum number of customers that the system can contain. Assume that at most the system can hold K customers and that any further arriving customers will refuse to enter the queue and will depart immediately without receiving service. This system is equivalent to a birth and death process with the following rates:

$$\lambda_k = \begin{cases} \lambda & k<K \\ 0 & k\geq K \end{cases} \tag{3.64}$$

$$\mu_k = \mu \quad k = 1,2,\ldots K \, . \tag{3.65}$$

Thus, as soon as the system fills up, the input is effectively turned off. The state-transition-rate diagram is depicted in Figure 3.9.

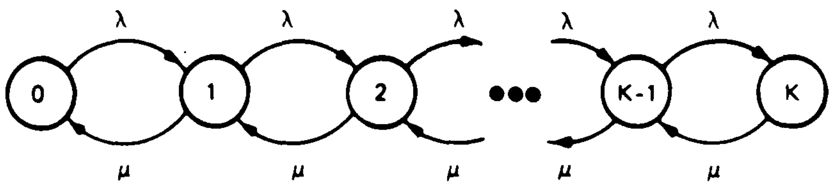


Figure 3.9 State-Transition-Rate Diagram for M/M/1/K system.

Obviously the equations are the same as for the M/M/1 case except that P_k =0 when k>K. Hence,

$$P_k = \begin{cases} (\lambda/\mu)^k \; P_0 & k \leq K \\ 0 & k > K \; . \end{cases} \qquad (3.66)$$

Solving for P_0 is somewhat more difficult, but again the conservation of probabilities equation is used.

$$P_0 + \sum_{k=1}^{K} (\lambda/\mu)^k \; P_0 = 1$$

$$P_0 \sum_{k=0}^{K} (\lambda/\mu)^k = 1$$

$$P_0 \left[\frac{1-(\lambda/\mu)^{K+1}}{1-(\lambda/\mu)}\right] = 1$$

$$P_0 = \frac{1-(\lambda/\mu)}{1-(\lambda/\mu)^{K+1}} \; . \qquad (3.67)$$

Hence,

$$P_k = \frac{[1-(\lambda/\mu)](\lambda/\mu)^k}{1-(\lambda/\mu)^{K+1}} \; . \qquad (3.68)$$

Since there is only a single server the utilization is

$$\rho = 1 - P_0 \; , \qquad (3.69)$$

and mean throughput is

$$T = \rho \; \mu \; . \qquad (3.70)$$

Another quantity of interest in this system is the probability that an arriving customer finds the system full, and therefore *leaves* without receiving service. This probability is

$$P_K = \frac{[1-(\lambda/\mu)](\lambda/\mu)^K}{1-(\lambda/\mu)^{K+1}} \; . \qquad (3.71)$$

3.8 M/M/1//M - Finite Customer Population - Single Server

This model is often referred to as the machine repair model. Consider a job shop which consists of M machines and one serviceman. Assume that the amount of time each machine runs before breaking down is exponentially distributed with mean 1/λ, and assume that the amount of time for the serviceman to repair a machine is exponentially distributed with mean 1/μ. The birth and death rates for such a system are:

$$\lambda_k = \begin{cases} (M-k)\lambda & k\leq M \\ 0 & k\geq M \end{cases} \qquad (3.72)$$

$$\mu_k = \mu \quad k = 1,2,3,\ldots \quad . \qquad (3.73)$$

Since there is only one serviceman the service rate is μ, regardless of the number of machines down. On the other hand, if k machines are not in use, then since the M-k machines in use each fail at a rate λ, it follows that $\lambda_k=(M-k)\lambda$. In the sense that a failing machine is regarded as an arrival and a repaired machine as a departure, the system represents a queueing system with a finite population. The state-transition-rate diagram is depicted in Figure 3.10.

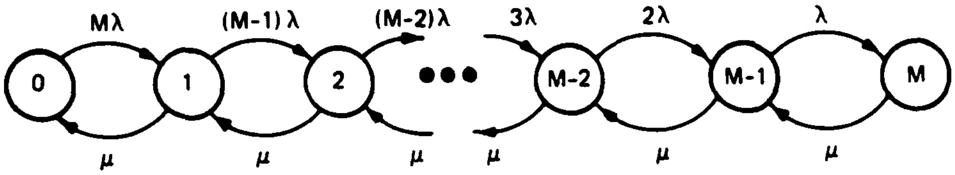


Figure 3.10 State-Transition-Rate Diagram for M/M/1//M System.

The resulting steady-state equations are:

state	rate out = rate in
0	$M\lambda P_0 = \mu P_1$
1	$[(M-1)\lambda+\mu]P_1 = M\lambda P_0+\mu P_2$
2	$[(M-2)\lambda+\mu]P_2 = (M-1)\lambda P_1+\mu P_3$
3	$[(M-3)\lambda+\mu]P_3 = (M-2)\lambda P_2+\mu P_4$
$k, k\geq 1$	$[(M-k)\lambda+\mu]P_k = (M+1-k)\lambda P_{k-1}+\mu P_{k+1}$. (3.74)

Solving these equation in terms of P_0 yields

$$P_1 = M\ (\lambda/\mu)\ P_0$$

$$P_2 = M(M-1)\ (\lambda/\mu)^2\ P_0$$

$$P_3 = M(M-1)(M-2)\ (\lambda/\mu)^3\ P_0$$

$$P_k = \begin{cases} [M!/(M-k)!]\ (\lambda/\mu)^k\ P_0 & k\leq M \\ 0 & k\geq M\ . \end{cases} \tag{3.75}$$

Solving for P_0 is the usual way results in:

$$P_0 = \Big[\sum_{k=0}^{M} [M!/(M-k)!]\ (\lambda/\mu)^k\Big]^{-1}\ . \tag{3.76}$$

Since there is only a single server the utilization is

$$\begin{aligned}\rho &= 1 - P_0 \\ &= 1 - \left[\sum_{k=0}^{M} [M!/(M-k)!]\ (\lambda/\mu)^k\right]^{-1} \\ &= \frac{\left[\sum_{k=0}^{M} [M!/(M-k)!]\ (\lambda/\mu)^k\right] - 1}{\sum_{k=0}^{M} [M!/(M-k)!]\ (\lambda/\mu)^k} \\ &= \frac{\sum_{k=1}^{M} [1/(M-k)!]\ (\lambda/\mu)^k}{\sum_{k=0}^{M} [1/(M-k)!]\ (\lambda/\mu)^k} . \qquad (3.77)\end{aligned}$$

Finally letting i=m-k and changing variables results in:

$$\rho = \frac{\sum_{i=1}^{M-1} (1/i!)\ (\lambda/\mu)^i}{\sum_{i=0}^{M} (1/i!)\ (\lambda/\mu)^i} . \qquad (3.78)$$

Similarly, the mean number of customers in the system is:

$$L = \frac{\sum_{k=1}^{M} k \; [M!/(M-k)!] \; (\lambda/\mu)^k}{\sum_{k=0}^{M} [M!/(M-k)!] \; (\lambda/\mu)^k} = M - \frac{\mu \rho}{\lambda} . \qquad (3.79)$$

The mean throughput is

$$T = \rho \mu . \qquad (3.80)$$

Again using Little's Law to find the response time yields:

$$R = \frac{L}{T} = \frac{M-(\lambda/\mu)\rho}{\rho \mu} = \frac{M}{\rho \mu} - \frac{1}{\mu} . \qquad (3.81)$$

3.9 Other Elementary Queueing System

Several other queueing systems can be modeled by judicious assignments of the rates λ_k and μ_k. For example, the following systems are solved in Kleinrock : M/M/m/m - M server loss system, M/M∞//M - finite customer population - infinite number of servers, M/M/m/K/M - finite population - m servers - finite storage, and other cases including discouraged arrivals [KLEI75].

CHAPTER 4

QUEUEING MODELS WITH GENERAL SERVICE OR ARRIVAL PATTERNS

4.1 The M/G/1 Queueing System

The M/G/1 model represents the contention for a single server under the assumption that the arrival process is Poisson. Thus, this model is more general than the M/M/1 in that there are no restrictions on the distribution of the service times. The difficulty in analyzing this model stems from the fact that the distribution of the service times is not memoryless. Information about when the service started assists in predicting when the service will be completed. Hence, the number of customers presently in the system is not enough information to predict the number of customers in the future. Therefore, the process can no longer be represented as a continuous-time Markov chain with the number of customers in the system serving as the state space.

However, if the system is observed only at departure instants the past history plays no roll in predicting the future. This is because service is just starting, and prior information cannot aid in predicting when it will be completed. The past history also cannot help with arrivals since the interarrival times have a negative exponential distribution and are therefore memoryless. Hence, if the system is observed only at departure instants (immediately after a departure) the system appears to be a Markov chain. Such a process is referred to as semi-Markov process with an embedded discrete-time Markov chain. That is, the behavior of the system at the departure instants can be described by a Markov chain. Fortunately, the solution at these embedded points happens also to provide the solution for all points in time [COHE69] [CINL75].

Recall from Chapter 3 that the limiting probabilities of a discrete-time Markov chain can be found from

$$V = V[P] \tag{4.1}$$

where $V = [P_0, P_1, P_2, \ldots]$ and [P] is the one-step transition matrix. The elements of [P] are the one-step transition probabilities:

$$P_{ij} = P[X_{n+1}=j \mid X_n=i] \ . \tag{4.2}$$

That is, P_{ij} is the conditional probability that the next state is j, given that the current state is i. Since the embedded process is obtained from the continuous process by observing the system immediately after a departure, it follows that

$$X_{n+1} = \begin{cases} X_n - 1 + A_{n+1} & \text{for } X_n \geq 1 \\ A_{n+1} & \text{for } X_n = 0, \end{cases} \tag{4.3}$$

where X_n is the number of customers in the system at the nth departure point and A_{n+1} is the number of customers who arrive during the service time of the (n+1)st customer. Thus, $j < i-1$ is an impossible situation whereas $j \geq i-1$ is possible for all values since any number of arrivals can occur during one service time. It follows that the form of the one-step transition matrix is:

$$[P] = \left| \begin{array}{cccccc} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ 0 & 0 & 0 & a_0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{array} \right| \tag{4.4}$$

where a_k = P[k arrivals during one service time]. For example $P_{i,i-1}$ is the probability that zero arrivals occur during one service period, and $P_{i,i}$ is the probability that one arrival occurs during the service period (the one arrival offsets the one departure). Also, note that the first two rows of this matrix are identical. This is because if a departing customer leaves an empty system, the state remains zero until an arrival occurs. The graphical form of the process is depicted in Figure 4.1. The labels on the arcs are probabilities.

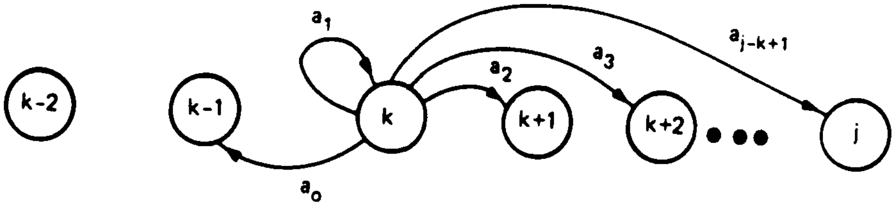


Figure 4.1 State-Transition-Probability Diagram for the M/G/1 System.

Now, since the arrival process is Poisson with rate λ, the conditional probability of k arrivals, given that the service time is τ, is

$$P[\text{k arrivals} \mid \text{the service time} = \tau] = \frac{e^{-\lambda\tau} \; (\lambda\tau)^k}{k!} \quad . \tag{4.5}$$

Thus, the unconditional probability of k arrivals is,

$$a_k = \int_0^\infty \frac{e^{-\lambda\tau} \; (\lambda\tau)^k}{k!} \, dB(\tau) \quad , \tag{4.6a}$$

or

$$a_k = \int_0^\infty \frac{e^{-\lambda\tau} \; (\lambda\tau)^k}{k!} \, b(\tau) \; d\tau \quad , \tag{4.6b}$$

where $B(\tau)$ is the probability distribution function (PDF) of the service times, and $b(\tau)$ the probability density (pdf) of the service times.

Returning to the problem of finding the limiting probabilities, it follows that the component form of Equation (4.1) for the M/G/1 case is

$$P_k = P_0 \; a_k + \sum_{i=1}^{k+1} P_i \; a_{k+1-i} \quad . \tag{4.7}$$

This equation can be solved by the method of probability generating functions. By definition the probability generating function is

$$F(z) = \sum_{k=0}^{\infty} f_k \; z^k \; . \tag{4.8}$$

The procedure is to use the probability generating function to transform Equation (4.02) into a function of z, and to then solve for

P(z). Once P(z) has been determined, the limiting probabilities can be found from the series expansion of P(z). More precisely, the coefficient of the z^i term is P_i. Multiplying both sides of Equation (4.7) by z^k and summing from k=0 to k=∞, results in

$$\sum_{k=0}^{\infty} P_k\ z^k = \sum_{k=0}^{\infty} P_0\ a_k\ z^k + \sum_{k=0}^{\infty} \sum_{i=1}^{k+1} P_i\ a_{k+1-i}\ z^k \quad . \qquad (4.9)$$

Interchanging the order of the double summation and simplifying results in :

$$P(z) = P_0\ a(z) + \sum_{i=1}^{\infty} \sum_{k=i-1}^{\infty} P_i\ a_{k+1-i}\ z^k \ ,$$

$$= P_0\ a(z) + \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} P_i\ a_j\ z^k$$

$$= P_0\ a(z) + z^{-1} \sum_{i=1}^{\infty} P_i\ z^i \sum_{j=0}^{\infty} a_j\ z^j$$

$$= P_0\ a(z) + z^{-1}\ [P(z)-P_0]\ a(z) \ . \qquad (4.10)$$

Solving this last equation for P(z), results in

$$P(z) = \frac{P_0\ a(z)\ [z-1]}{z-a(z)} \quad . \qquad (4.11)$$

From the definition of the probability generating function, a(z) is

$$a(z) = \sum_{k=0}^{\infty} \Big[\int_0^{\infty} \frac{e^{-\lambda\tau}\ (\lambda\tau)^k}{k!}\ b(\tau)\ d\tau \Big] z^k \quad . \qquad (4.12)$$

Interchanging the order of the summation and the integration, results in:

$$a(z) = \int_0^\infty \sum_{k=0}^\infty \frac{e^{-\lambda\tau} \; (\lambda\tau z)^k}{k!} b(\tau) \; d(\tau)$$

$$= \int_0^\infty e^{-(\lambda-\lambda z)\tau} \; b(\tau) \; d(\tau) \; . \qquad (4.13)$$

The last equation is the same as the Laplace transform of $b(\tau)$ except that the parameter s has been replaced by $\lambda-\lambda z$. Let $b^*(s)$ denote the Laplace transform of $b(\tau)$, then $a(z) = b^*(\lambda-\lambda z)$. Substituting the result into Equation (4.11) results ir

$$P(z) = \frac{P_0 \; b^*(\lambda-\lambda z) \; [z-1]}{z - b^*(\lambda-\lambda z)} \cdot \qquad (4.14)$$

P_0 can be determine from this last equation by taking the limit of both sides as z approaches one. The limit of $P(z)$ as z approaches one is :

$$\lim_{z\to 1} \sum_{k=0}^\infty P_k \; z^k = \sum_{k=0}^\infty \lim_{z\to 1} P_k \; z^k = \sum_{k=0}^\infty P_k = 1 \quad . \qquad (4.15)$$

In addition the limit of $b^*(\lambda-\lambda z)$ as z approaches one is $b^*(0) = 1$. That is, $b(\tau)$ is a density function, and

$$\int_0^\infty b(\tau) \; d\tau = 1 \quad . \qquad (4.16)$$

It follows that the limit as z approaches one of Equation (4.14) is indeterminate and l'Hopital's rule must be used. The derivative of $b^*(\lambda-\lambda z)$ evaluated at $z=1$ is

$$\frac{db^*(\lambda-\lambda z)}{dz}\bigg|_{z=1} = \frac{d}{dz}\left[\int_0^\infty e^{-(\lambda-\lambda z)}\ b(\tau)\ d\tau\right]_{z=1}$$

$$= \int_0^\infty \frac{d}{dz}\left[e^{-(\lambda-\lambda z)\tau}\right]_{z=1} b(\tau)\ d\tau$$

$$= \lambda \int_0^\infty \tau\ b(\tau)\ d\tau$$

$$= \lambda E[S]\ , \qquad (4.17)$$

where E[S] is the average or expected value of the service time. Using this and applying l'Hopital's rule results in

$$P_0 = 1 - \lambda E[S]\ . \qquad (4.18)$$

The utilization is therefore

$$\rho = 1 - P_0 = \lambda E[S]\ . \qquad (4.19)$$

Substituting this into Equation (4.11) yields the Pollaczek-Khinchin (P-K) transform equation

$$P(z) = \frac{(1-\lambda E[S])\ b^*(\lambda-\lambda z)\ (z-1)}{1 - b^*(\lambda-\lambda z)}\ , \qquad (4.20a)$$

or equivalently

$$P(z) = \frac{(1-\rho)\ b^*(\lambda-\lambda z)\ (z-1)}{1 - b^*(\lambda-\lambda z)}\ . \qquad (4.20b)$$

As derived in Chapter 3 the mean value of P_k equals the derivative of P(z) evaluated at z=1. After using l'Hopital's rule twice the result is :

$$L = \frac{\lambda^2 E[S^2]}{2(1-\lambda E[S])} + \lambda E[S] \ . \qquad (4.21)$$

The appearance of the second moment in this equation comes from the fact that:

$$\left.\frac{d^2 b^*(\lambda-\lambda z)}{dz^2}\right|_{z=1} = \lambda^2 E[S^2] \ . \qquad (4.22)$$

The mean response time can be calculated from Little's law. More precisely,

$$R = \frac{L}{T} = \frac{L}{\lambda} = \frac{\lambda E[S^2]}{2(1-\lambda E[S])} + E[S] \ . \qquad (4.23)$$

The mean waiting time is obviously $W_q = R - E[S]$. Hence,

$$W_q = \frac{\lambda E[S^2]}{2(1-\lambda E[S])} \ . \qquad (4.24)$$

The mean number of customers in the queue can be determined from W_q. That is

$$L_q = TW_q = \lambda W_q$$

$$= \frac{\lambda^2 E[S^2]}{2(1-\lambda E[S])} \ . \qquad (4.25)$$

As an example consider the M/M/1 system. The density function of for the service times is

$$b(\tau) = \mu e^{-\mu\tau} \ , \qquad (4.26)$$

and

$$b^*(s) = \frac{\mu}{s+\mu} \ , \qquad (4.27)$$

$$b^{*}(\lambda-\lambda z) = \frac{\mu}{\lambda-\lambda z+\mu} \cdot \qquad (4.28)$$

Substituting into Equation (4.20b) results in

$$P(z) = \frac{1-\rho}{1-\rho z} \cdot \qquad (4.29)$$

P(z) can be expanded into positive powers of z by simply dividing 1-ρ by 1-ρz. The result is

$$\frac{1-\rho}{1-\rho z} = (1-\rho) \; [\; 1 + \rho z + \rho^2 z^2 + \rho^3 z^3 + \cdots \;] \; . \qquad (4.30)$$

Thus,

$$P_k = \rho^k(1-\rho) \qquad (4.31)$$

which is the same as before.

4.1.1 Comments on the Steady-State Solution and that of the Embedded Markov Chain

Early in this chapter it was stated that the steady-state solution for all time and that of the embedded Markov chain at the departure instants were the same. Unfortunately, there is no simply way to prove this statement. It was, however, shown that P_0 of the embedded process was 1-λE[S], which agrees with the result in Chapter 1, which is valid for all work conservative single queueing systems. Hence, P_0 is the same for both processes. Also, since P_0 is the long-run proportion of the time that the system is idle, the expected values of the normalized idle and busy periods are the same.

The approach taken here of simply stating that the solution of both processes is the same, is that taken by most texts on queueing theory [KLEI75] [ALLE78] [ROSS80] [KOBA81] [HAYE84]. The reader is advised to beware of short simple proofs claiming to prove that both processes have the same solution. In particular the proofs in Gross and Cooper are incomplete [GROS74] [COOP84]. Both prove simply the probability that an arriving customer finds k customers in the system is equal to the probability that a departing customers leave k in the system. As pointed out in Ross and Kleinrock this is true, not only for the M/G/1 system, but also for the M/M/1, G/M/1 and G/G/1 systems [ROSS80] [KLEI75]. Furthermore, it is proven by a counter example in Ross that an arriving or departing customer does not necessarily see time averages. That is, the probability that an arriving customer finds k in the system is not necessarily the same as P_k. However, both Ross and Kleinrock state, without proof or references, that if the arrival process is Poisson then an arriving customer sees time averages.

Additional comments on this subject are contained in another section in this chapter.

4.1.2 M/D/1 - Poisson Inputs - Constant Service time

As a second example of a M/G/1 system, consider the case in which the arrival process is Poisson with mean rate λ and the service time constant with rate μ. Since the service time is constant

$$E[S] = 1/\mu \, , \tag{4.32}$$

and

$$E[S^2] = (E[S])^2 = 1/\mu^2 \, . \tag{4.33}$$

Substituting these results into equations (4.21), (4.23), (4.24) and (4.25) yields:

$$L = \frac{(\lambda/\mu)^2}{2(1-\lambda/\mu)} + \frac{\lambda}{\mu}$$

$$= \frac{\rho^2}{2(1-\rho)} + \rho$$

$$= \frac{\rho}{(1-\rho)} - \frac{\rho^2}{2(1-\rho)} , \qquad (4.34)$$

$$R = \frac{\lambda/\mu^2}{2(1-\lambda/\mu)} + \frac{1}{\mu}$$

$$= \frac{\rho}{2\mu(1-\rho)} + \frac{1}{\mu}$$

$$= \frac{1}{\mu(1-\rho)} - \frac{\rho}{2\mu(1-\rho)} , \qquad (4.35)$$

$$L_q = \frac{(\lambda/\mu)^2}{2(1-\lambda/\mu)} = \frac{\rho^2}{2(1-\rho)} , \qquad (4.36)$$

$$W_q = \frac{\lambda/\mu^2}{2(1-\lambda/\mu)} = \frac{\rho}{2\mu(1-\rho)} . \qquad (4.37)$$

The mean performance equations for the M/D/1 system are compared to those for the M/M/1 system in Table 4.1. They are also compared graphically in Figures 4.2, 4.3, 4.4, and 4.5. Observe that L, L_q, R, and W_q are all less for the M/D/1 system. This is because equations (4.21) and (4.23)-(4.25) are directly proportional to the second moment

Parameter	M/D/1	M/M/1
ρ	$\lambda E[S]$	$\lambda E[S]$
L	$\frac{\rho}{(1-\rho)} - \frac{\rho^2}{2(1-\rho)}$	$\frac{\rho}{(1-\rho)}$
R	$\frac{1}{\mu(1-\rho)} - \frac{\rho}{2\mu(1-\rho)}$	$\frac{1}{\mu(1-\rho)}$
L_q	$\frac{\rho^2}{2(1-\rho)}$	$\frac{\rho^2}{(1-\rho)}$
W_q	$\frac{\rho}{2\mu(1-\rho)}$	$\frac{\rho}{\mu(1-\rho)}$

Table 4.1 Comparison of M/D/1 and M/M/1 Equations.

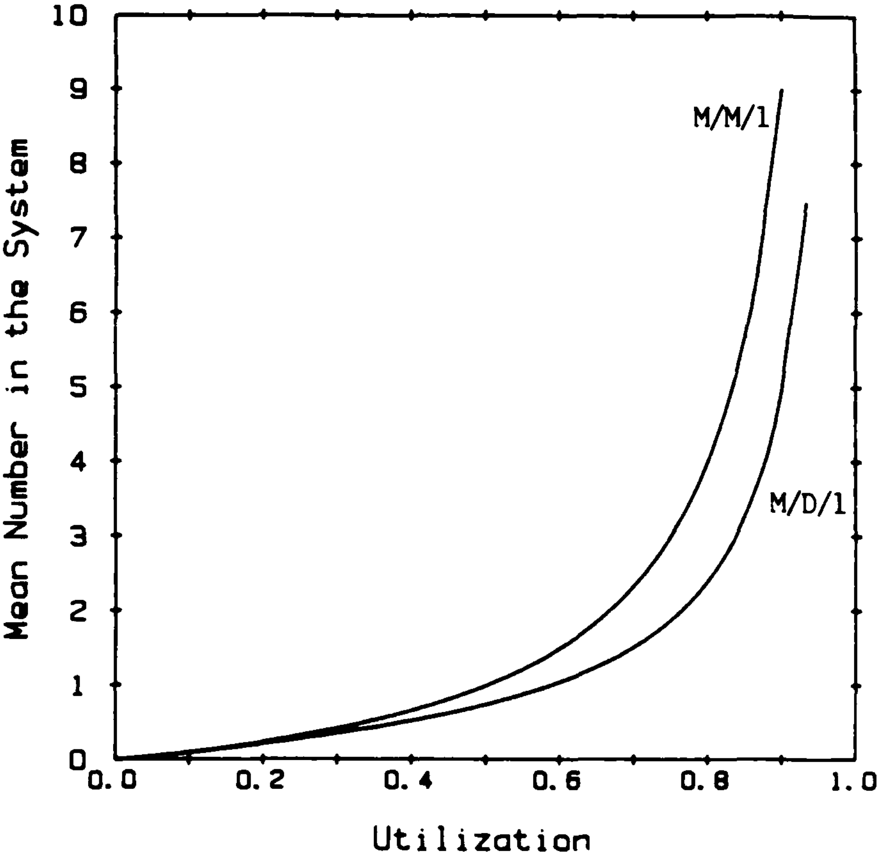


Figure 4.2 Curves of L Versus ρ for M/D/1 and M/M/1.

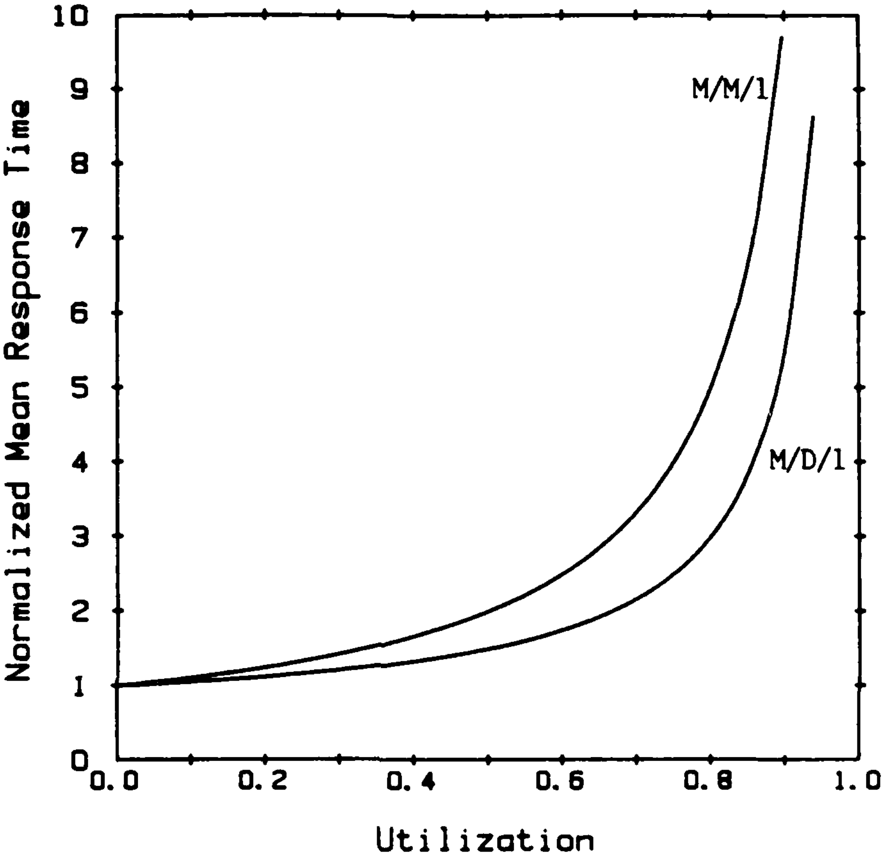


Figure 4.3 Curves of R Versus ρ for M/D/1 and M/M/1.

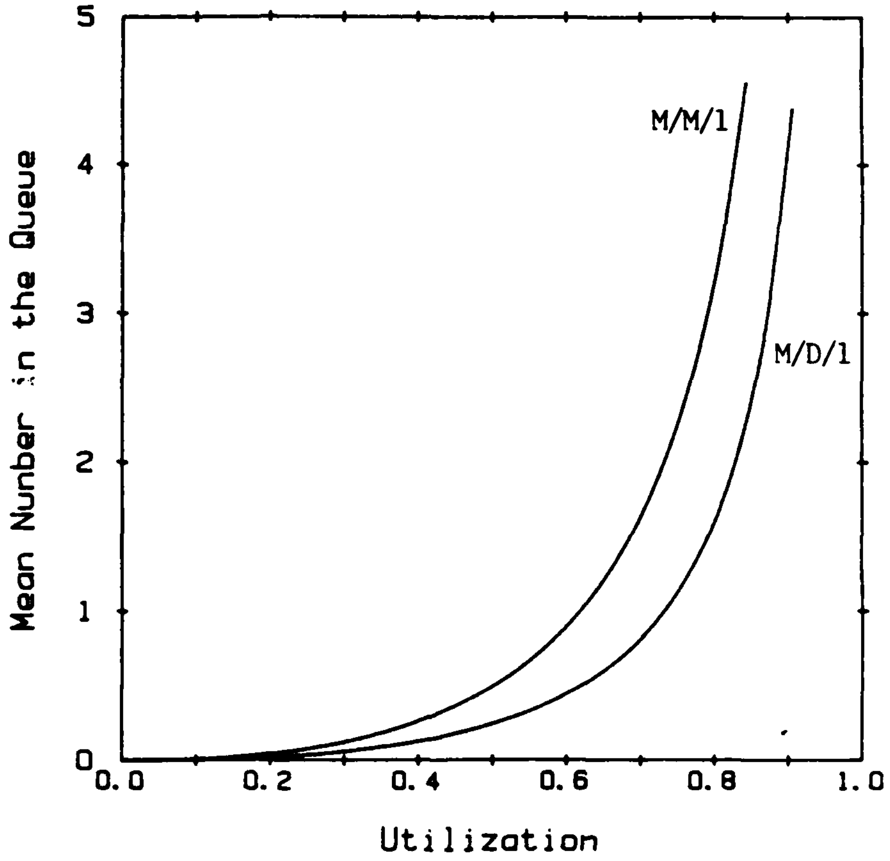


Figure 4.4 Curves of L_q Versus ρ for M/D/1 and M/M/1.

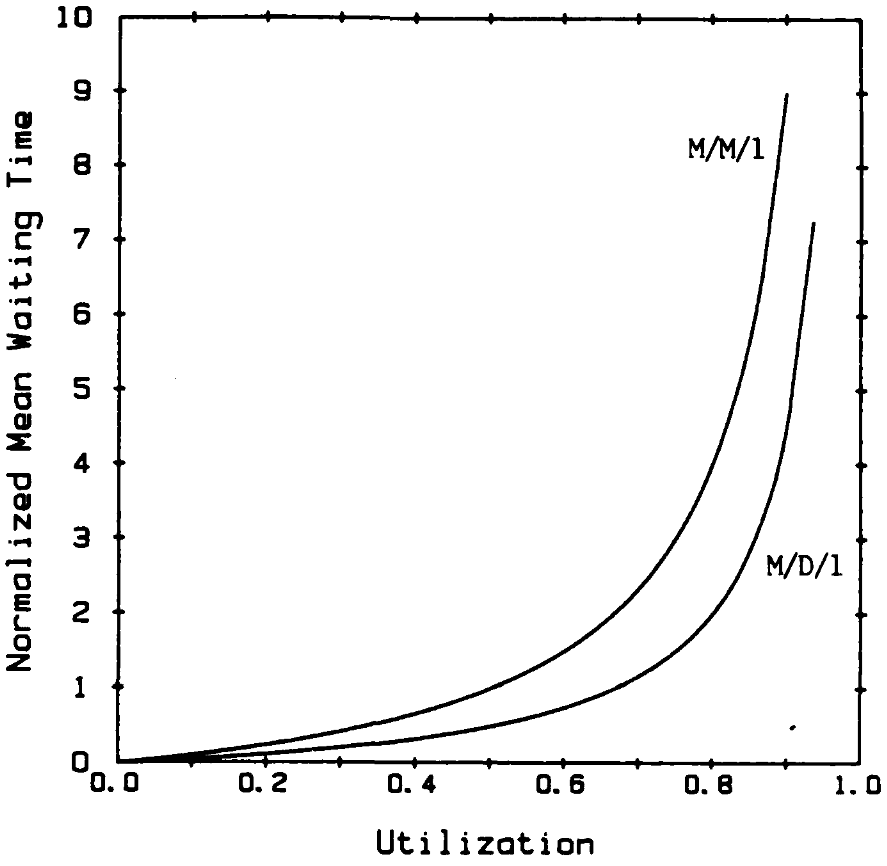


Figure 4.5 Curves of W_q Versus ρ for M/D/1 and M/M/1.

of the service time. That is, for a fixed mean value, as the second moment or variance increases so does L, L_q, R, and W_q. Also note that L_q and W_q are exactly one-half of that for the M/M/1 system. This results from the fact that the second moment of service time for the M/M/1 is $2/\mu^2$, which is exactly twice that of the M/D/1 system.

Although they will not be derived here, the probability generating function, customer distribution and variance for the M/D/1 system are [LAVE83]:

$$P(z) = \frac{(1-\rho)\ (1-z)}{1 - ze^{\rho(1-z)}} , \tag{4.38}$$

$$P_k = (1-\rho) \sum_{j=0}^{k} (-1)^{k-j} \frac{(j\rho)^{k-j-1}(j\rho+k-j)e^{j\rho}}{(k-j)!} , \tag{4.39}$$

$$\mathrm{Var}[N] = \rho + \frac{3\rho^2+2\rho^3}{6(1-\rho)} + \frac{\rho^4}{4(1-\rho)} . \tag{4.40}$$

Notice that these equations (and their derivations) are much more complicated than the corresponding expressions for the M/M/1 system. That is, the fact that service time is constant drastically complicates the analysis rather than simplifying it!

4.1.3 M/G/1 Nonpreemptive Priority

While this might not seem to be the appropriate place to discuss the priority service discipline, all the results here apply also to the M/M/1 nonpreemptive priority queue, and furthermore these are the only known results. A priority queueing system is one in which customers are grouped into classes and then given priority according to their class. Although there are several service disciplines based on priority, only the nonpreemptive discipline will be discussed here. While it would be nice to have an explicit expression for the probability distribution of customers or a transform expression, no one has derived such an expression. Thus, the following analysis is concerned with determining the mean values of the performance parameters.

It is assumed that the customers are divided into n classes numbered 1 to n, and that the lower the priority number the higher the priority. That is, customers in priority class i are given preference over customers in class j, if i<j. Customers within a priority class are served with respect to that class by the FCFS rule.

It is also assumed that the *arrival process* is Poisson. More precisely, class i customers arrive from a Poisson source at an average rate of λ_i. Hence the combined arrival process is Poisson with rate λ, where $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Each class of customers may have its own general service time distribution. Hence, the combined PDF of the service time is given by

$$B(\tau) = \frac{\lambda_1}{\lambda} B(\tau_1) + \frac{\lambda_2}{\lambda} B(\tau_2) + \cdots + \frac{\lambda_n}{\lambda} B(\tau_n) \ , \qquad (4.41)$$

where

$B(\tau_i)$ = PDF of the service time for class i customers,

and

$\frac{\lambda_i}{\lambda}$ = Probability the customer receiving service belongs to class i.

It follows that the expected value and second moment of the service time are respectively:

$$E[S] = \frac{\lambda_1}{\lambda} E[S_1] + \frac{\lambda_2}{\lambda} E[S_2] + \cdots + \frac{\lambda_n}{\lambda} E[S_n] \ , \qquad (4.42)$$

$$E[S^2] = \frac{\lambda_1}{\lambda} E[S_1{}^2] + \frac{\lambda_2}{\lambda} E[S_2{}^2] + \cdots + \frac{\lambda_n}{\lambda} E[S_n{}^2] \ . \qquad (4.43)$$

Since on the average λ_i customers arrive per second and these customers require an average of $E[S_i]$ seconds of service, then $\lambda_i E[S_i]$ is the percentage of time the server is serving class i customers. Therefore,

$$\rho = \rho_1 + \rho_2 + \cdots + \rho_n,$$

where

$$\rho_i = \lambda_i E[S_i] \ . \qquad (4.44)$$

Now suppose that a customer of priority i arrives at the system at time t_0 and starts to receive service a time t_1. His waiting time is thus $\tau_q = t_1 - t_0$. At t_0 let there be k_j ($j=1,2,\ldots,i$) customers of class j ahead of the arriving customers, and let there be either one or no customers in service at t_0. Also let $k_j{}'$ ($j=1,2,\ldots,i-1$) represent the number of class j customers that arrive during τ_q, and hence receive service before the customer who arrived at t_0. Now let

τ_j = total time required to service the k_j customers,

$\tau_j{}'$ = total time required to service the $k_j{}'$ customers,

and

τ_0 = the time required to finish serving the customer in service at t_0.

It follows that

$$\tau_q = \sum_{j=1}^{i-1} \tau_j' + \sum_{j=1}^{i} \tau_j + \tau_0 , \qquad (4.45)$$

and

$$W_{qi} = E[\tau_q] = \sum_{j=1}^{i-1} E[\tau_j'] + \sum_{j=1}^{i} E[\tau_j] + E[\tau_0] , \qquad (4.46)$$

where W_{qi} is the mean waiting time of a class i customer.

Since k_j and S_j are independent random variable it is easily seen that

$$E[\tau_j] = E[k_j]\ E[S_j]. \qquad (4.47)$$

Utilizing Little's law yields

$$E[k_j] = \lambda_j\ W_{qj} . \qquad (4.48)$$

Hence,

$$\begin{aligned} E[\tau_j] &= \lambda_j E[S_j]\ W_{qj} \\ &= \rho_j\ W_{qj}. \end{aligned} \qquad (4.48)$$

Similarly,

$$E[\tau_j'] = \rho_j\ W_{qi}. \qquad (4.50)$$

Substituting these results into the previous equation for W_{qi} yields

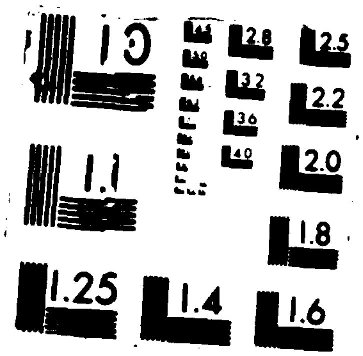
$$W_{qi} = W_{qi} \sum_{j=1}^{i-1} \rho_j + \sum_{j=1}^{i} \rho_j\ W_{qj} + E[\tau_0] , \qquad (4.51)$$

or

$$W_{qi} = \frac{\sum_{j=1}^{i} \rho_j\ W_{qj} + E[\tau_0]}{1 - \sigma_{i-1}} , \qquad (4.52)$$

AD-A182 994 NETWORKS OF MARKOVIAN QUEUES(U) ARMY ENGINEER WATERWAYS 2/4
EXPERIMENT STATION VICKSBURG MS INSTRUMENTATION
SERVICES DIV R A FRANCO MAY 87 WES/TR/O-87-2

UNCLASSIFIED F/G 12/3 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

where

$$\sigma_{i-1} = \sum_{j=1}^{i-1} \rho_j \ .$$

By induction on i one obtains

$$W_{qi} = \frac{E[\tau_0]}{(1-\sigma_{i-1})\ (1-\sigma_i)} \ . \tag{4.53}$$

In order to determine $E[\tau_0]$ assume that n=1, and that the arrival and service processes are the same as the earlier combined processes. Hence, there is only one class of customers and they are served in FCFS order. Therefore, W_{q1} equals W_q for a M/G/1 system, and

$$W_{q1} = \frac{E[\tau_0]}{1-\rho} = \frac{\lambda E[S^2]}{2(1-\rho)} \ . \tag{4.54}$$

Solving for $E[\tau_0]$ yields

$$E[\tau_0] = \frac{\lambda E[S^2]}{2} \ . \tag{4.55}$$

Finally, substituting this equation into the last expression for W_{qi} results in

$$W_{qi} = \frac{\lambda E[S^2]}{2(1-\sigma_{i-1})(1-\sigma_i)} \ . \tag{4.56}$$

The mean values of the other parameters follow directly from the last expression. More precisely,

$$R_i = E[S_i] + W_{qi}, \qquad (4.57)$$

$$L_{qi} = \lambda_i W_{qi}, \qquad (4.58)$$

$$L_i = \lambda_i R_i. \qquad (4.59)$$

The only other equation that has been derived for the nonpreemptive queueing discipline is the variance of the response time. The equation is given without proof :

$$\mathrm{Var}[W] = \mathrm{var}[S_i] + \frac{\lambda E[S^3]}{3[1-\sigma_{i-1}]^2[1-\rho]}$$

$$+ \frac{\lambda E[S^2]\ 2\left[\sum_{j=1}^{i} \lambda_j E[S_j{}^2] - \lambda E[S^2]\right]}{4[1-\sigma_{i-1}]^2[1-\rho]}$$

$$+ \frac{\lambda E[S^2] \sum_{j=1}^{i-1} \lambda_j E[S_j{}^2]}{2[1-\sigma_{i-1}]^3[1-\rho]} \qquad (4.60)$$

[LAVE83]. Unfortunately, these are all the equations that have been derived for the nonpreemptive priority queueing discipline. They are far short of what one would need in order to determine the queue size or buffer size so that overflow does not occur.

The behavior of priority queues is illustrated graphically in Figures 4.6, and 4.7. Figure 4.6 is plot of L_1, L_2, L_3, and L_T (total) for the M/M/1 priority system. It is assumed that all arrival rates and service time distributions are the same for all three classes. It

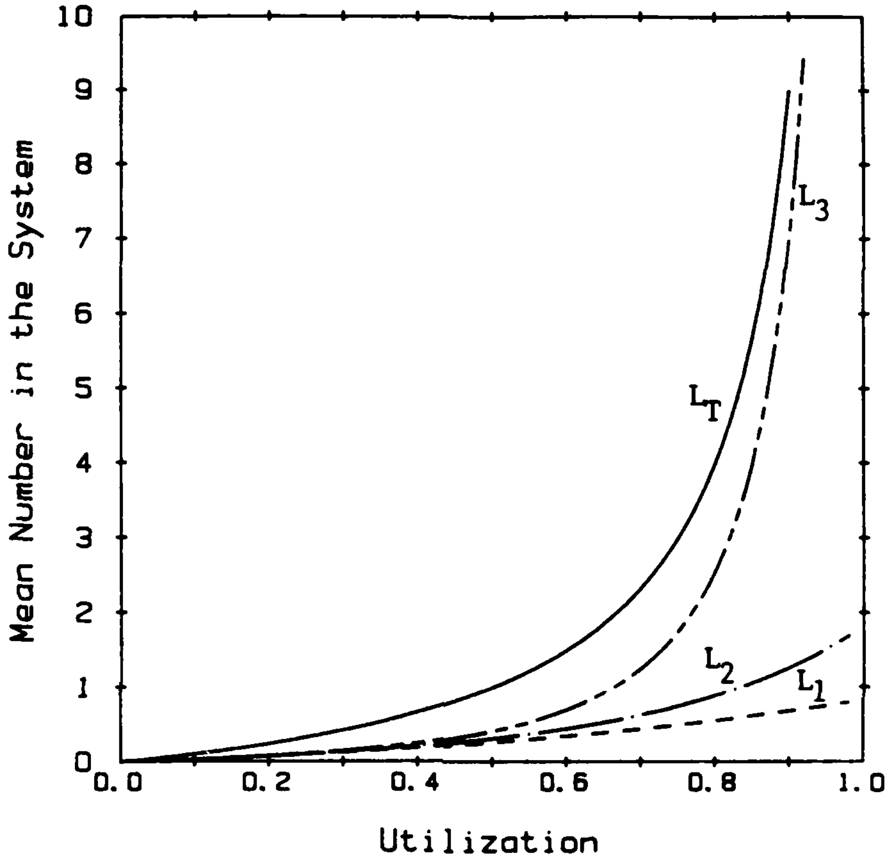


Figure 4.6 Curves of L_1, L_2, L_3 and L_T for a M/M/1 Priority System with Three Customer Classes.

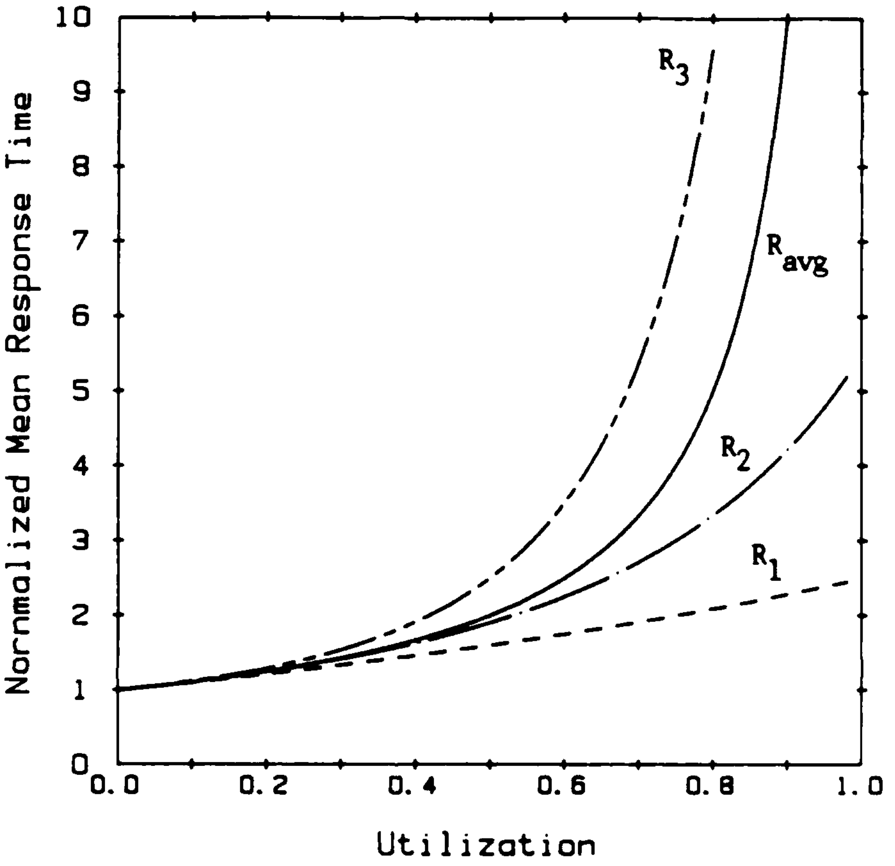


Figure 4.7 Curves of R_1, R_2, R_3 and R_{avg} for a M/M/1 Priority System with Three Customer Classes.

should not be surprising that L_T is the same as the M/M/1 system with one class and an arrival rate of $\lambda_1+\lambda_2+\lambda_3$. Figure 4.7 is a plot of the normalized mean response time (E[S]=1) for the same system. Note that in both figures the service discipline has the most effect on the class with the lowest priority (i=3). Figures 4.8 and 4.9 are similar to Figures 4.6 and 4.7 except that the service time is deterministic.

4.1.4 Comments on the M/G/m Queueing Model

The M/G/m model represents the contention for m identical servers that operate independently in parallel under the assumption that the arrival process is Poisson. Thus, this model is a generalization of the M/M/m model. Although this model is often encountered in practice, analytical results have not been obtained for it. The primary reason for this is that it is not a semi-Markov process. More precisely the number of customers in the system at the departure instants is not enough information to predict future behavior. Information concerning the amount of service received by customers at the other servers is relevant. Hence, the system does not possess an embedded Markov chain at the departure instants.

4.2 The G/M/1 Queueing System

The G/M/1 queueing model represents the contention for a single server under the assumption that the interarrival times have a general distribution and that the service times have an exponential distribution. Compared to the M/G/1 model few analytical results are available.

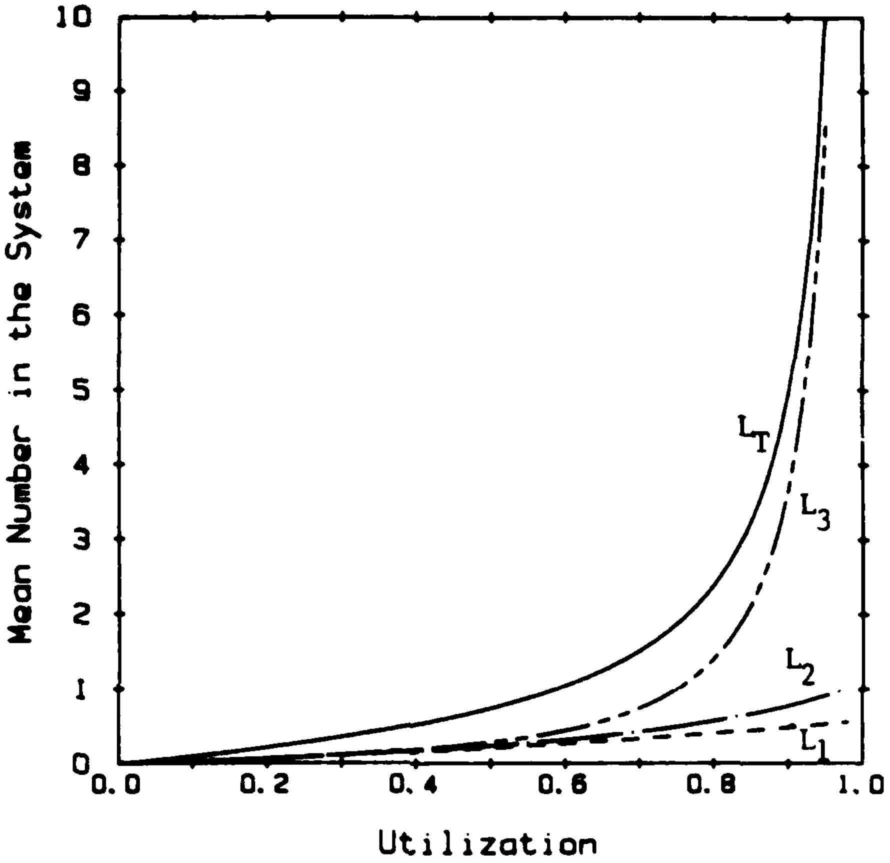


Figure 4.8 Curves of L_1, L_2, L_3 and L_T for a M/D/1 Priority System with Three Customer Classes.

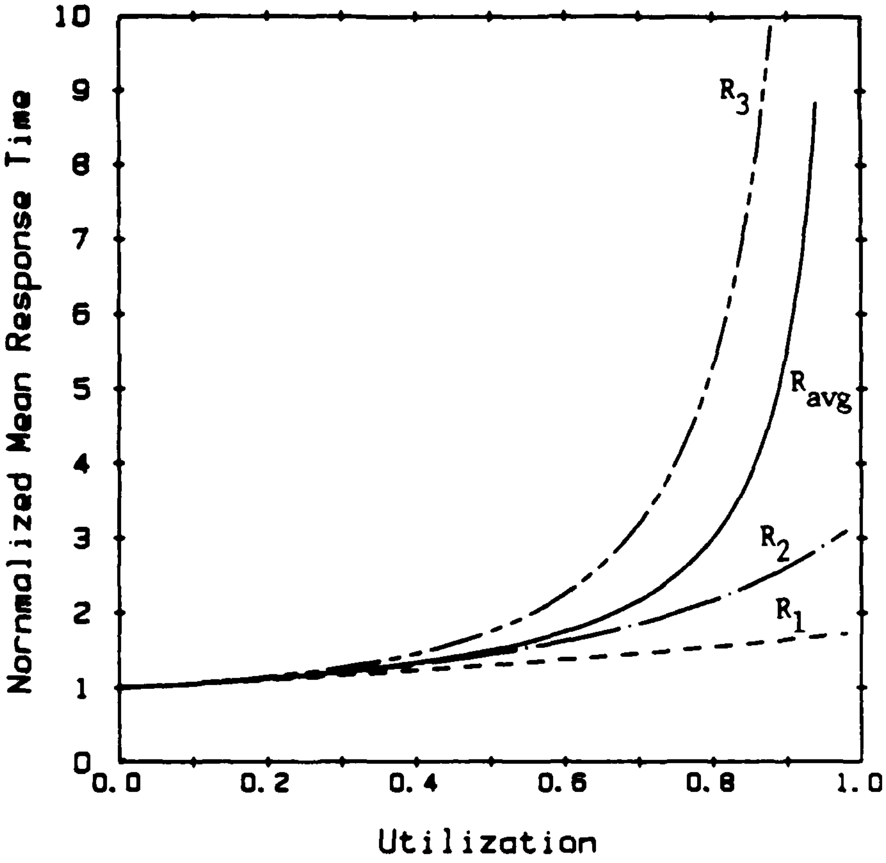


Figure 4.9 Curves of R_1, R_2, R_3 and R_{avg} for a M/D/1 Priority System with Three Customer Classes.

Since the arrival process is not memoryless, the process is not Markovian. However, if the system is observed only just before an arrival, it appears to be a discrete-time Markov chain. Hence, the process is a semi-Markov process with an embedded chain. Unfortunately, the solution at these embedded points is not the solution for all points in time [COHE69] [GROS74] [CINL75]. However, it is possible to determine the mean values of the performance parameters from the solution at these points.

Since the solution at these embedded points is not the steady-state solution for all time, the symbol P_k will not be used to represent the limiting probabilities. Instead the symbol π_k will be used. Thus, the discrete-time limiting probability equation becomes

$$[\pi_0,\pi_1, \ldots] = [\pi_0,\pi_1, \ldots] [P] , \qquad (4.61)$$

where [P] is the one-step transition matrix.

The state-transition-probability diagram for the G/M/1 model is depicted in Figure 4.10. Note that a transition from state i to j where

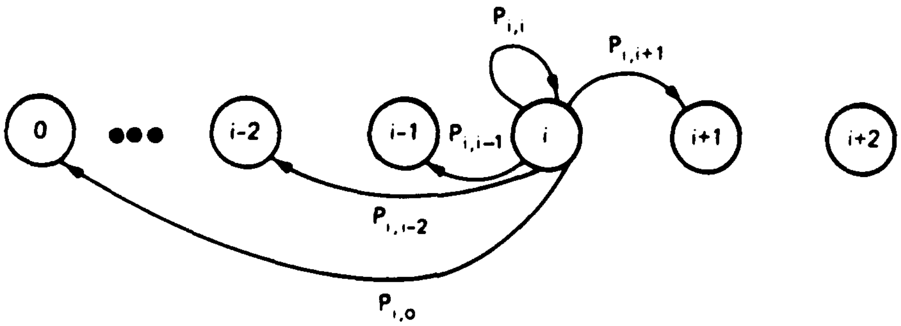


Figure 4.10 State-Transition-Probability Diagram for the G/M/1 System.

$j > i+1$ is an impossible situation since only one arrival can occur during an interarrival period. On the other hand, up to $i+1$ departures can occur during an interarrival period, therefore all transitions to state j where $0 \leq j \leq i+1$ are possible.

In order to determine the form of the one-step transition matrix, note that the following relationship exist between states X_{n+1} and X_n

$$X_{n+1} = X_n + 1 - B_n, \qquad (4.62)$$

where B_n denotes the number of departures between the nth and (n+1)th arrival. Thus, the form of the one-step transition matrix is

$$P = \left| \begin{matrix} 1-b_0 & b_0 & 0 & 0 & 0 & \cdots \\ 1-\sum\limits_{m=0}^{1} b_m & b_1 & b_0 & 0 & 0 & \cdots \\ 1-\sum\limits_{m=0}^{2} b_m & b_2 & b_1 & b_0 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{matrix} \right| \qquad (4.63)$$

where b_n = P[n services during an interarrival period]. Recall that the elements of the one-step transition matrix are P_{ij}, and note that $n=i+1-j$. Also note that the case $j=0$ is treated separately because, if $j=0$, it is not sufficient to say that $i+1-j$ customers were served during an interarrival period. That is, they could have been served in less time.

In order to calculate b_n, recall that if the service process is Markovian, the service time is exponentially distributed. Hence, as long as there are customers to be served, the number of services in any length of time t is a Poisson random variable with mean μt. Thus, if A(t) is the PDF of the interarrival times, then by conditioning on the time between successive arrivals:

$$b_n = \int_0^\infty \frac{e^{-\mu t}(\mu t)^n}{n\,!}\, dA(t) \qquad 0\leq n<i+1 \; . \qquad (4.65)$$

Hence, in component form the limiting probability equation becomes:

$$\pi_k = \sum_{n=0}^{\infty} \pi_{k+n-1}\, b_n \qquad k>1, \qquad (4.66)$$

or

$$\pi_k = \sum_{n=0}^{\infty} \pi_{k+n-1} \int_0^\infty \frac{e^{-\mu t}(\mu t)^n}{n!}\, dA(t) \qquad k>1. \qquad (4.67)$$

The π_0 equation has not been included since it contains no new information, it is redundant. The value π_0 can be determined from the fact that the limiting probabilities must sum to one.

Unfortunately, there is no easy way to solve this last equation, however, it has been proven that if $\mu/\lambda < 1$ (the necessary condition for the limiting probabilities to exist), then the form of the solution is

$$\pi_k = c\beta^k, \qquad (4.68)$$

where β = a number between 0 and 1,

and c = a constant which will be determined

[TAKA62]. Substituting this into the last equation leads to :

$$c\beta^k = \sum_{n=0}^{\infty} c\beta^{k+n-1} \int_0^{\infty} \frac{e^{-\mu t}(\mu t)^n}{n!} \, dA(t) \qquad k\geq 1$$

$$= c\beta^{k-1} \int_0^{\infty} e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\beta\mu t)^n}{n!} \, dA(t)$$

$$= c\beta^{k-1} \int_0^{\infty} e^{-\mu t(1-\beta)} \, dA(t) \ . \qquad (4.69)$$

Hence,

$$\beta = \int_0^{\infty} e^{-\mu(1-\beta)t} \, dA(t) \ . \qquad (4.70)$$

Observe that this last equation is just the Laplace transform of $a(t)$ evaluated at $\mu(1-\beta)$. That is

$$\beta = a^*(s)\Big|_{s=\mu(1-\beta)} = a^*(\mu-\mu\beta). \qquad (4.71)$$

The exact value of β usually can only be determined by numerical analysis (such as Newton's method).

The constant c can be determined from

$$\sum_{k=0}^{\infty} \pi_k = 1, \qquad (4.72)$$

which implies

$$c \sum_{k=0}^{\infty} \beta^k = 1, \tag{4.73}$$

or

$$c = (1-\beta). \tag{4.74}$$

Hence,

$$\pi_k = \beta^k \ (1-\beta),$$

where

$$\beta = a^*(\mu-\mu x). \tag{4.75}$$

It is important to emphasize that π_k is not the steady-state probability of k customers in the system. It is probability that an arriving customer finds k customers in the system. Now, if an arriving customer find k customer in the system, it follows that his expected response time is (k+1)/μ (this is true regardless of how much service the current customer has already received since the service distribution is Markovian and thus memoryless). Hence, the mean response time can be determined by conditioning on the number in the system when a customer arrives. That is

$$R = \sum_{k=0}^{\infty} E[\text{time in system} \mid \text{arrival sees k}] \ \beta^k(1-\beta)$$

$$= \sum_{k=0}^{\infty} \frac{k+1}{\mu} \ \beta^k(1-\beta)$$

$$= \frac{1}{\mu(1-\beta)}, \tag{4.76}$$

where the last step follows from the identity

$$\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2} \; . \qquad (4.77)$$

The mean values of the other performance parameters easily follow:

$$L = \lambda R = \frac{\lambda}{\mu(1-\beta)} \; , \qquad (4.78)$$

$$W_q = R - \frac{1}{\mu} = \frac{\beta}{\mu(1-\beta)} \; , \qquad (4.79)$$

$$L_q = \lambda W_q = \frac{\lambda\beta}{\mu(1-\beta)} \; . \qquad (4.80)$$

4.2.1 Comments on the G/M/m Queueing Model

The G/M/m queueing model represents the contention for m identical severs that operate independently and in parallel under the assumption that the interarrival times have an exponential distribution. Similar to the G/M/1 system if this system is observed at the arrival instants then it appears to be a Markov chain. Hence it is possible to calculate the probability that an arriving customer finds k customers in the system. However, no one has been able to derive explicit expressions for any performance parameter. The details of the analysis are significantly more complicated than those for the G/M/1 system and will not be given here. The interested reader is referred to Gross and Kleinrock [GROS74] [KLEI75].

4.3 Comments on the Solution of a Semi-Markov Process and the General Time Process

It should be obvious that a Markov process is also a semi-Markov process. Therefore, if relationships could be developed that relate the solution of a semi-Markov process to that of its general time process, then the M/M/1, M/G/1, and G/M/1 systems could all be analyzed as semi-Markov processes. Indeed such relationships have been developed [FABE61] [CINL69] [CINL75]. However, each case must be treated as a separate problem. The development of the relationships depends upon renewal theory, and requires far too much background material to be presented here. The interested reader is referred to the references by Cinlar, who developed much of the theory. The analysis is by no means simple. The author spends an entire chapter (chapter 10) developing the relationships for the M/G/1 and G/M/1 cases [CINL75]. The primary results are that the solution of the semi-Markov process for the M/G/1 case is the same as the general time solution, whereas the general time solution for the G/M/1 case is

$$P_k = \begin{cases} 1 - \dfrac{\lambda}{\mu} & k=0 \\ \dfrac{\lambda}{\mu}\,\beta^{k-1}(1-\beta) & k\geq 1. \end{cases} \qquad (4.81)$$

It is somewhat ironic that the proof for the M/G/1 case is considerably more complicated than that for the G/M/1 case.

4.4 Comments on the G/G/1 Queueing Model

The G/G/1 queueing model represents contention for a single server under the conditions that both the interarrival times and the service times have general distributions. Clearly, this case includes the

M/M/1, M/G/1, and G/M/1 cases. Therefore, if a solution could be obtained in terms of the system parameters it would be valid for the other cases. Unfortunately no one has derived such solution.

The difficulty in analyzing the model stems from the fact that neither the arrival nor service process is memoryless. Hence, it is not possible to define a Markov or semi-Markov process where the state of the system represents the number of customer. However, it is possible to define a semi-Markov process where the state represents the amount of unfinished work in the system. The embedded process is obtained by looking at the system only at customer arrival instants. Thus, the unfinished work at these points is the same as the customers response time, R.

The details of the analysis will not be covered here, but some comments on the analysis and form of the solution will be discussed. The semi-Markov process is a discrete-time continuous-state process. Note that this is our first encounter with a process in which the state space is continuous. The key point here is that in order to obtain a solution for the waiting time, complex variable theory and spectral factoring must be used. The procedure involves a certain amount of trial and error. Unfortunately, the spectral factoring procedure destroys all traces of the system parameters. That is, although it may be possible to find a solution, it will not be in terms of any of the system parameters. However, assuming that a solution can be found, the expected values of the other performance parameters can be determined from the waiting time.

Although it is not yet possible to find an expression, in terms of the system parameter, it is possible to derive an expression for the upper bound of L and R. The results are:

$$L \leq \rho + \frac{\lambda^2(\mathrm{Var}[\eta]+\mathrm{Var}[S])}{2(1-\rho)} , \tag{4.82}$$

$$R \leq E[S] + \frac{\lambda(\mathrm{Var}[\eta]+\mathrm{Var}[S])}{2(1-\rho)} , \tag{4.83}$$

where

η = random variable representing the interarrival time, [LAVE83].

4.4.1 Comments on the G/G/m Queueing System

All of the comments on the G/G/1 system carry over to the G/G/m case. However, the likelihood of solving the integral equation by special factoring is considerable less than the G/G/1 case (usually impossible [KLEI76]). Bounds on L and R have been derived [LAVE83]:

$$L \leq m\rho + \frac{\lambda^2[\mathrm{Var}[\eta]+(\mathrm{Var}[S]/m)]}{2(1-\rho)}, \tag{4.84}$$

$$W \leq E[S] + \frac{\lambda[\mathrm{Var}[\eta]+(\mathrm{Var}[S]/m)]}{2(1-\rho)} . \tag{4.85}$$

4.5 Concluding Remarks

The primary purpose of this chapter was to develop the elementary results for the M/G/1 and G/M/1 systems. It is not possible to cover all of the details of these two systems in a single chapter. Indeed entire books have been devoted to covering these two systems. Perhaps the most comprehensive reference is 'The Single Server Queue' by Cohen [COHE69]. This reference contains over 600 pages!

CHAPTER 5

INTRODUCTION TO MARKOVIAN QUEUEING NETWORKS

5.1 Introduction to Queueing Networks

A queueing network is a collection of one or more service centers. Only networks of Markovian queues will be considered in this chapter. A Markovian network is one in which all arrivals from outside the network are Poisson processes and all service times are exponentially distributed. The purpose of a queueing network is to predict the performance of a physical system in which there is contention for resources. The resources are represented by the servers in the network. Queueing networks are usually classified as being either open or closed An open network in depicted in Figure 5.1. A customer enters one of the service centers from outside the system, waits for a server to become free, receives service, and departs the service center. Upon departing from the service center the customer, according to fixed routing probabilities, either enters another service center, reenters the same service center or exits the system. Open networks are used to model systems in which the number of customers competing for resources can be potentially unlimited. A closed queueing network is shown in Figure 5.2. In a closed queueing network the number of customers in the system is always constant. After a customer completes service at a service center, he either enters another service center or reenters the same service center. Closed networks are used to model systems in which a fixed number of customers contend for the resources.

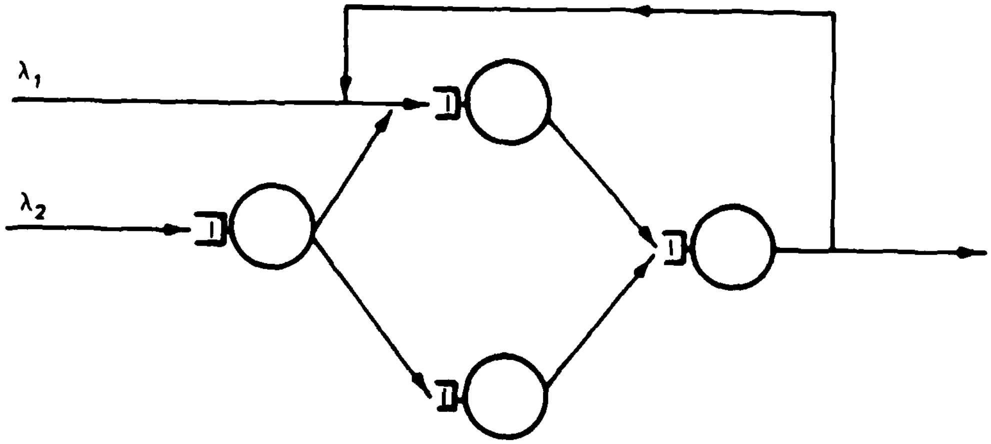


Figure 5.1 An Open Queueing Network.

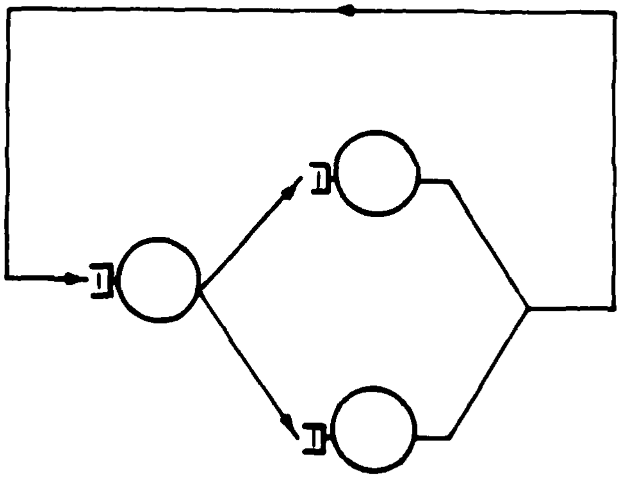


Figure 5.2 A Closed Queueing Network.

5.2 Burke's Theorem

Before surging into queueing networks first consider the simple two node (service center) network of Figure 5.3. Assume that both service

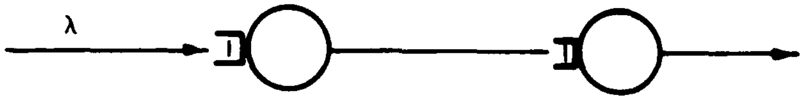


Figure 5.3 A Simple Tandem Queueing Network.

centers contain a single server and that the service times are exponentially distributed with mean $1/\mu_1$ at node one and $1/\mu_2$ at node two. Also assume that the arrival process to node one is Poisson with rate λ. Thus the first node is exactly an M/M/1 queue. In order to analyze the second node the arrival process feeding it must be calculated. Clearly, this is the departure process of node one. Let D(t) denote the PDF of the interdeparture time between customers leaving node one. When a customer departs node one, either a second customer immediately starts service or the queue is empty. If the queue is empty, then the time until the next customer departs is the sum of two independent random variables: the first being the time until a new customer arrives and the second his service time. The density function of the sum of two independent random variables is the convolution of the individual density functions [KLEI75]. Therefore, it is easier to work with the density function and Laplace transforms and then convert back. Let d(t) denote the density function of the interdeparture process at service center one and $d^*(s)$ its Laplace transform. The conditional Laplace transform densities are :

$$d^*(S) \,|\, \text{node one empty} = \frac{\lambda}{S+\lambda} \; \frac{\mu}{S+\mu} \tag{5.1}$$

and

$$d^*(S) \,|\, \text{node one nonempty} = \frac{\mu}{S+\mu} \tag{5.2}$$

where the subscripts have been suppressed since all variables pertain to service center one. The probability that an M/M/1 queue is nonempty was calculated earlier and is λ/μ. Therefore the unconditional Laplace transform density is

$$d^*(S) = \left[1 - \frac{\lambda}{\mu}\right] \left[\frac{\lambda}{S+\lambda} \; \frac{\mu}{S+\mu}\right] + \left[\frac{\lambda}{\mu} \; \frac{\mu}{S+\mu}\right] = \frac{\lambda}{S+\lambda} \, , \tag{5.3}$$

and

$$d(t) = \lambda e^{-\lambda t}. \tag{5.4}$$

Hence, the interdeparture PDF is

$$D(t) = 1 - e^{\lambda t}. \tag{5.5}$$

Thus the departure process of an M/M/1 process is exactly the same as the arrival process! This startling result is usually referred to as Burke's theorem [BURK56]. He also proved that the same was true for an M/M/m queue.

In view of Burke's theorem, service center two is also an M/M/1 queue with mean arrival rate λ and can be analyzed independently of node one. It follows that the joint probability of node one containing k_1 customers and node two containing k_2 customers is

$$P(k_1,k_2) = P_1(k_1)\ P_2(k_2)$$

$$= \left[(\lambda/\mu_1)^{k_1} P_1(0)\right] \left[(\lambda/\mu_2)^{k_2} P_2(0)\right]$$

$$= (\lambda/\mu_1)^{k_1} [1-(\lambda/\mu_1)] \quad (\lambda/\mu_2)^{k_2} [1-(\lambda/\mu_2)]. \quad (5.6)$$

5.3 Open Queueing Networks and Jackson's Product Form Theorem

Shortly after Burke published his work Jackson proved that more general networks can be analyzed in a similar manner [JACK57]. The type of networks he studied is depicted in Figure 5.1 . It consists of N interconnected nodes. Node i in the network contains m_i identical servers. The service time of a customer visiting node i is exponentially distributed with mean $1/\mu_i$. A customer after receiving service at node i is routed to node j according to probability r_{ij}, or he reenters node i according to probability r_{ii}, or he exits the network according to probability r_{i0}. In addition to receiving customers from other nodes, node i may receives customers from a Poisson process outside the network at mean rate λ_i.

Let γ_i denote the total mean arrival rate at node i (arrivals from outside the network plus those from other nodes inside the network). Since the expected value of the sum of several random variables is the sum of the individual expected values (irrespective of dependencies involved), it follows that at steady-state

$$\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j \, r_{ji} \quad , \qquad (5.7)$$

where $\gamma_j \, r_{ji}$ is the mean rate from node j to i. Hence a set of N linear, simultaneous equations can be written from which the mean arrival rate at each node can be determined. It also follows that the mean throughput, $T_i = \gamma_i$.

Jackson proved for this class of networks that the joint probability distribution is

$$P(k_1,k_2,\ldots,k_N) = P_1(k_1)\ P_2(k_2) \cdots P_N(k_N) \tag{5.8}$$

where

$$P_i(k_i) = \begin{cases} P_i(0)\ (\gamma_i/\mu_i)^{k_i}\ /\ k_i! & (k_i = 0,1,\ldots,m_i) \\ \\ P_i(0)\ (\gamma_i/\mu_i)^{k_i}\ /\ (m_i!\ m_i^{k_i-m_i}) & (k_i = m_i,m_i+1,\ldots)\ . \end{cases}$$

Note that $P_i(k_i)$ is the same equation that was derived earlier for an M/M/m queue except γ_i has replaced λ_i. The result is known as Jackson's product form theorem.

The proof closely parallels that of the birth and death model in chapter three. In fact a network of Markovian queues is a multidimensional birth and death model. Recall that for the one-dimensional birth and death model the probability of zero births in an infinitesimal interval h is $1-\lambda h+o(h)$. It follows that for a network of N nodes the probability of zero births in h is

$$[1-\lambda_1 h+o(h)]\ [1-\lambda_2 h+o(h)]\ \ldots\ [1-\lambda_N h+o(h)] = 1 - \sum_{i=1}^{N} \lambda_i h + o(h)\ . \tag{5.9}$$

Similarly, the probability of zero deaths at node i is

$$1 - \mu_i(k_i)\ h + o(h)\ , \tag{5.10}$$

where

$$\mu_i(k_i) = \begin{cases} k_i\ \mu_i & \text{for } k_i \leq m_i \\ m_i\ \mu_i & \text{for } k_i \geq m_i \end{cases}$$

μ_i = mean service rate at queue i when k_i=1

k_i = the number of customers at service center i

m_i = the number of servers at center i.

The probability of zero deaths in the networks is

$$1 - \sum_{i=1}^{N} \mu_i(k_i)\ h + o(h)\ , \tag{5.11}$$

and the joint probability of zero births and zero deaths is

$$1 - \sum_{i=1}^{N} \lambda_i\ h - \sum_{i=1}^{N} \mu_i(k_i)\ h + o(h)\ . \tag{5.12}$$

By considering all the ways in which a network can reach state $(k_1,k_2,\ldots,k_N)$ it turns out that

$$\begin{aligned} P(k_1,\ldots,k_N)(t+h) = {} & \left[1 - \sum_{i=1}^{N} \lambda_i\ h - \sum_{i=1}^{N} \mu_i(k_i)\ h\ \right] P(k_1,\ldots,k_N)(t) \\ & + \sum_{i=1}^{N} \lambda_i\ h\ P(k_1,\ldots,k_i-1,\ldots,k_N)(t) \\ & + \sum_{i=1}^{N} \mu_i(k_i+1)\ h\ r_{i0}\ P(k_1,\ldots,k_i+1,\ldots,k_N)(t) \\ & + \sum_{i=1}^{N} \sum_{j=1}^{N} \mu_j(k_j+1-\delta_{ij})\ h\ r_{ji}\ P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)(t) \\ & + o(h)\ , \end{aligned} \tag{5.13}$$

where

$$\delta_{ij} = \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases} .$$

The first term on the right has already been explained. The second term on the right is the probability that the network is in state $(k_1,\ldots,k_i-1,\ldots,k_N)$, and an arrival occurs at node i in time h. State $(k_1,\ldots,k_i-1,\ldots,k_N)$ indicates that service center i has one less customer than state $(k_1,\ldots,k_i,\ldots.k_N)$. The third term is the probability of the network being in state $(k_1,\ldots,k_i+1,\ldots,k_N)$, and a departure occurs at service center i, and the departing customer exits the network. The fourth term is the probability that the network is in state $(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)$, and a departure occurs at service center j, and the departing customer goes to service center i. The δ_{ij} terms allow for the possibility that a departing customer reenters the same service center.

Following the usual procedure of subtracting $P(k_1,\ldots,k_N)(t)$ from both sides, dividing by h and taking the limit as h approaches zero, one obtains a set of differential equations. A set of steady-state equations is then obtained by taking the limit as t approaches infinity. The resulting steady-state equations are :

$$\left[\sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N} \mu_i(k_i) \right] P(k_1,\ldots,k_N) = \sum_{i=1}^{N} \lambda_i \; P(k_1,\ldots,k_i-1,\ldots,k_N)$$

$$+ \sum_{i=1}^{N} \mu_i(k_i+1) \; r_{i0} \; P(k_1,\ldots,k_i+1,\ldots,k_N)$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{N}\mu_j(k_j+1-\delta_{ij})\; r_{ji}\; P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)\cdot \tag{5.14}$$

Jackson did not derive the solution from this set of equations as was done in the one-dimensional case. He assumed the solution and then verified that his assumption was correct by substituting it into these equations. The following relations are easily seen from the defining equation for $P(k_1,k_2,\ldots,k_N)$:

$$\frac{P(k_1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_i,\ldots,k_N)} = \frac{\mu_i(k_i)}{\gamma_i} \tag{5.15}$$

$$\frac{P(k_1,\ldots,k_i+1,\ldots,k_N)}{P(k_1,\ldots,k_i,\ldots,k_N)} = \frac{\gamma_i}{\mu_i(k_i+1)} \tag{5.16}$$

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{\gamma_j\;\mu_i(k_i)}{\gamma_i\;\mu_j(k_j+1-\delta_{ij})}\;. \tag{5.17}$$

Dividing both sides of the steady-state equation by $P(k_1,\ldots,k_N)$ results in :

$$\sum_{i=1}^{N}\lambda_i + \sum_{i=1}^{N}\mu_i(k_i) = \sum_{i=1}^{N}\lambda_i\;\mu_i(k_i)\;/\;\gamma_i + \sum_{i=1}^{N}\gamma_i\; r_{i0}$$

$$+\sum_{j=1}^{N}\sum_{i=1}^{N}\gamma_i\;\mu_i(k_i)\; r_{ji}\;/\;\gamma_i\;. \tag{5.18}$$

Substituting $\lambda_i = \gamma_i + \sum_{j=1}^{N}\gamma_j\; r_{ji}$ into the first summation on the right and adding the result to the double summation yields:

$$\sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N} \mu_i(k_i) = \sum_{i=1}^{N} \gamma_i \; r_{i0} + \sum_{i=1}^{N} \mu_i(k_i) \; . \qquad (5.19)$$

Finally, substituting $r_{i0} = 1 - \sum_{j=1}^{N} r_{ij}$ and $\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j \; r_{ji}$ into the last equation results in both sides becoming identical, and the proof is completed.

Jackson's product form theorem states that once the mean arrival rates have been determined, each service center can be analyzed as an independent M/M/m queue. As in the case of the M/M/m queue the service discipline or order in which customers are served is unimportant as long as it is work-conservative. The results of Jackson' theorem can best be illustrated by an example. Consider the problem of finding the distribution of customers in the network in Figure 5.4. Assume that both service centers have a single server. The following are the steady-state equations for all states with two or fewer customers:

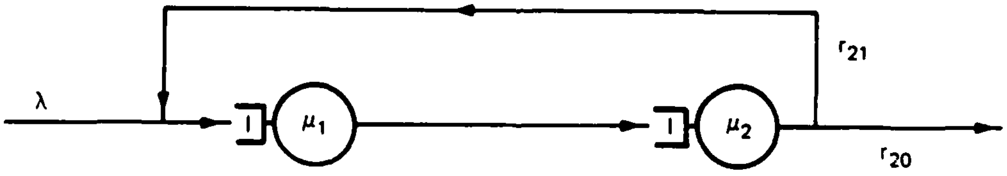


Figure 5.4 An Open Queueing Network with Feedback.

$\lambda\ P(0,0) = r_{2,0}\ \mu_2\ P(0,1)$

$(\lambda+\mu_1)\ P(1,0) = \lambda\ P(0,0) + r_{2,1}\ \mu_2\ P(0,1) + r_{2,0}\ \mu_2\ P(1,1)$

$(\lambda+\mu_2)\ P(0,1) = \mu_1\ P(1,0) + r_{2,0}\ \mu_2\ P(0,2)$

$(\lambda+\mu_1)\ P(2,0) = \lambda\ P(1,0) + r_{2,1}\ \mu_2\ P(1,1) + r_{2,0}\ P(2,1)$

$(\lambda+\mu_2)\ P(0,2) = \mu_1\ P(1,1) + r_{2,0}\ \mu_2\ P(0,3)$

$(\lambda+\mu_1+\mu_2)P(1,1) = \lambda\ P(0,1) + \mu_1\ P(2,0) + r_{2,1}\ \mu_2\ P(0,2)$
$\qquad + r_{2,0}\ \mu_2\ P(1,2).$

Notice that there are six equations and nine unknowns. No matter how many and what set of equations are written out there will always be more unknowns than equations. Thus, there is no way to solve this set of equations without some form of guessing.

According to Jackson's theorem the solution is

$$P(k_1,k_2) = (\gamma_1/\mu_1)^{k_1}\ P_1(0)\ (\gamma_2/\mu_2)^{k_2}\ P_2(0)\ ,$$

where γ_1 and γ_2 are the mean arrival rates. The equations for γ_1 and γ_2 are:

$$\gamma_1 = \lambda + \gamma_2\ r_{2,1}$$
$$\gamma_2 = \gamma_1\ .$$

Solving these equations results in $\gamma_1 = \gamma_2 = \lambda/r_{2,0}$. Which results in

$$P(k_1,k_2) = (\lambda/r_{2,0})^{k_1+k_2}\ (1/\mu_1)^{k_1}\ (1/\mu_2)^{k_2}\ P(0,0)\ .$$

It is easy to verify that this satisfies the steady-state equations, and thus is the solution.

5.3.1 Open Networks with Feedback

Jackson's product form theorem is not surprising for networks without feedback. If Poisson processes are joined or split the resulting processes are Poisson. Using this fact and Burke's theorem it can easily be shown that for a network without feedback the arrival process at node i is Poisson with mean rate γ_i.

The problem with feedback is that it can be proved that the arrival processes at service centers in a feedback loop are not Poisson [LEMO77]. This fact can be illustrated best by an example. Again using Figure 5.4 assume that customers arrive from outside the network from a Poisson source at an rate of one customer per hour, and that the mean service times at both service centers is exponentially distributed with a mean of 1 msec. Also, assume that the output from the second service center is fed back to first with probability r_{21}= 0.999. With this extreme set of parameters the output of the first service center tends to be in bursts. A typical output sequence is shown in Figure 5.5. The input to the second service center does not have independent increments, and therefore is not Poisson [ROSS80]. At present there is no explanation or conjecture as to why Jackson's theorem is valid for networks with feedback.

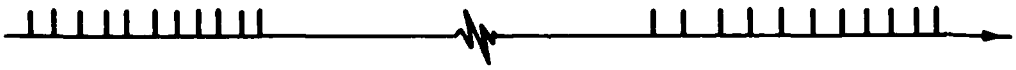


Figure 5.5 A Non-Poisson Input Sequence.

5.2.2 Local Balance

Although it was not necessary it turned out that steady-state equations are not only balanced, but that :

$$\mu_i(k_i)\ P(k_1,\ldots,k_N) = \lambda_i\ P(k_1,\ldots,k_i-1,\ldots,k_N) + \sum_{j=1}^{N} \mu_j(k_j+1-\delta_{ij})\ P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N), \tag{5.20}$$

and

$$\sum_{i=1}^{N} \lambda_i\ P(k_1,\ldots,k_N) = \sum_{i=1}^{N} \mu_i(k_i+1)\ r_{i0}\ P(k_1,\ldots,k_i+1,\ldots,k_N)\ . \tag{5.21}$$

These equations are called local balance equations as opposed to the steady-state equations which are often referred to as global balance equations. Local balance states that the rate of flow out of a network state due to a customer departing a queue is *equal to the rate of flow* into the state due to a customer arriving at the queue. Clearly the sum of the local balance equations are the global balance equations. Hence the solution to the local balance equations satisfies the global balance equations. Local balance was discovered by Whittle, and is also referred to as independent balance [WHIT68] [WHIT69]. As an example of local balance and its use, again consider the network in Figure 5.4 . The following are local balance equations that correspond to the global balance equations given earlier:

$$\lambda\ P(0,0) = r_{2,0}\ \mu_2\ P(0,1)$$

$$\lambda\ P(1,0) = r_{2,0}\ \mu_2\ P(1,1)$$

$$\mu_1\ P(1,0) = \lambda\ P(0,0) + r_{2,1}\ P(0,1)$$

$$\lambda\ P(0,1) = r_{2,0}\ \mu_2\ P(0,2)$$

$$\mu_2\ P(0,1) = \mu_1\ P(1,0)$$

$$\lambda\ P(2,0) = r_{2,0}\ \mu_2\ P(2,1)$$

$$\mu_1\ P(2,0) = \lambda\ P(1,0) + r_{2,1}\ \mu_2\ P(1,1)$$

$$\lambda\ P(0,2) = r_{2,0}\ \mu_2\ P(0,3)$$

$$\mu_2\ P(0,2) = \mu_1\ P(1,1)$$

$$\mu_1\ P(1,1) = \lambda\ P(0,1) + r_{2,1}\ \mu_2\ P(0,2)$$

$$\mu_2\ P(1,1) = \mu_1\ P(2,0)\ .$$

There are 12 local balance equations and only nine unknowns, therefore some of the equations are redundant. The following is a subset of the local balance equations:

$$\lambda\ P(0,0) = r_{2,0}\ \mu_2\ P(0,1)$$

$$\lambda\ P(0,1) = r_{2,0}\ \mu_2\ P(0,2)$$

$$\lambda\ P(1,0) = r_{2,0}\ \mu_2\ P(1,1)$$

$$\mu_1\ P(2,0) = \lambda\ P(1,0) + r_{2,1}\ \mu_2\ P(1,1)$$

$$\mu_2\ P(0,2) = \mu_1\ P(1,1)$$

$$\lambda\ P(2,0) = r_{2,0}\ \mu_2\ P(2,1)$$

$$\lambda\ P(1,1) = r_{2,0}\ \mu_2\ P(1,2)$$

$$\lambda\ P(0,2) = r_{2,0}\ \mu_2\ P(0,3)\ .$$

Solving these equations in terms of P(0,0) results in :

$$P(0,1) = (\lambda/r_{2,0})\ (1/\mu_2)\ P(0,0)$$

$$P(1,0) = (\lambda/r_{2,0})\ (1/\mu_1)\ P(0,0)$$

$$P(1,1) = (\lambda/r_{2,0})^2 \; (1/\mu_1) \; (1/\mu_2) \; P(0,0)$$
$$P(2,0) = (\lambda/r_{2,0})^2 \; (1/\mu_1)^2 \; P(0,0)$$
$$P(0,2) = (\lambda/r_{2,0})^2 \; (1/\mu_2)^2 \; P(0,0)$$
$$P(2,1) = (\lambda/r_{2,0})^3 \; (1/\mu_1)^2 \; (1/\mu_2) \; P(0,0)$$
$$P(1,2) = (\lambda/r_{2,0})^3 \; (1/\mu_1) \; (1/\mu_2)^2 \; P(0,0)$$
$$P(0,3) = (\lambda/r_{2,0})^3 \; (1/\mu_2)^3 \; P(0,0) \; .$$

The form of the solution is :

$$P(k_1,k_2) = (\lambda/r_{2,0})^{k_1+k_2} \; (1/\mu_1)^{k_1} \; (1/\mu_2)^{k_2} \; P(0,0) \; .$$

Which agrees with the solution obtained earlier. This is the only known example of a queueing network (with two or more service centers) being solved by local balance.

It is important to note that to show that local balance exist Jackson's theorem was used. It is also important to emphasize that local balance is not necessary for global balance. That is although all queueing networks discussed thus far have local balance others do not. Local balance is an important tool that can be used to help determine if advanced queueing networks have a product form solution. In other words one can assume local balance and if the local balance equations are consistent an answer can be obtained. The answer can then be verified and used to prove the assumption was true.

5.3.3 An Application of an Open Queueing Network

An example of an open queueing network model is shown in Figure 5.6 [FERR78]. The model represents a mainframe computer that consists of an input/output processor (IOP), a central processing unit (CPU), a

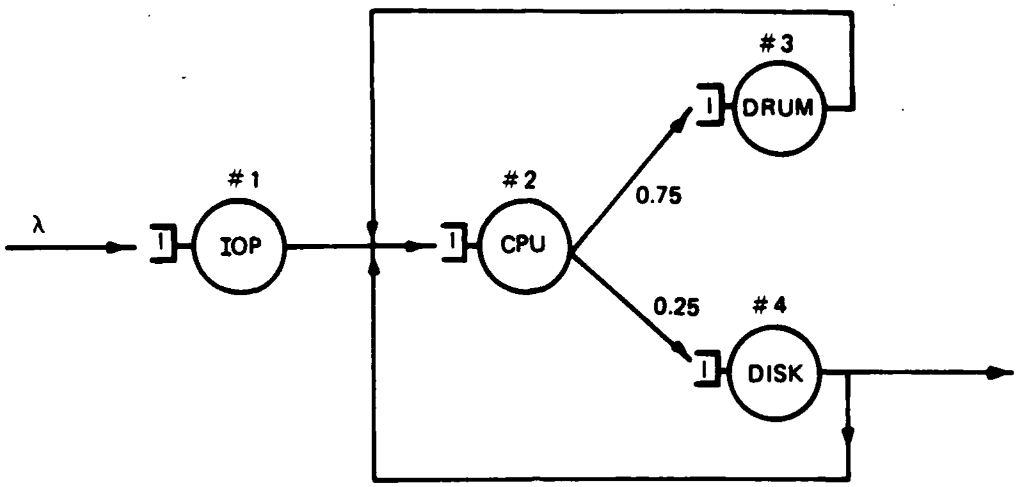


Figure 5.6 An Open Queueing Network Model of a Computer.

Parameter name	Symbol	Value
Mean arrival rate	λ	0.7 jobs/s
Mean input service time	$1/\mu_1$	500ms
Mean uninterrupted CPU time	$1/\mu_2$	30ms
Mean drum service time	$1/\mu_3$	20ms
Mean disk service time	$1/\mu_4$	80ms
IOP to CPU probability	r_{12}	1.00
CPU to drum probability	r_{23}	0.75
CPU to disk probability	r_{24}	0.25
Drum to CPU probability	r_{32}	1.00
Disk to CPU probability	r_{42}	0.90
Disk to out probability	r_{40}	0.10

Table 5.1 Parameter of the Model in Figure 5.6.

drum processor, and a disk processor. Each unit is assumed to consist of a single server, and the service discipline is assumed to be FCFS. Parameters for the model are given in table 5.1. The input process to the computer is assumed to be Poisson. All jobs are assumed to be statistically identical, and all service times are assumed to be exponentially distributed. The service time at the IOP accounts for the time to input and load a job into primary memory. After the job has been loaded it waits its turn to be processed by the CPU. The mean service time at the CPU represents the mean time before the job needs a drum or disk operation. Because these operations are slow compared to the speed of the CPU, the CPU releases the job to the drum or disk processor and starts on another one. The drum and disk service times represent typical demands made by jobs. It is assumed that all jobs that require drum operations will also require more CPU time before completion, therefore they are routed back to the CPU. On the other hand it is assumed that only 90 percent of the jobs that require disk service will require more CPU time. Two simplications have been made that violate a real system. First the model does not take in to account contention for memory, and second a job usually terminates at the IOP so that its output can reach the external world. These simplications are necessary in order to get an answer.

Since there is only one entry point and one exit point the mean throughput rate must equal the mean arrival rate. This is not, however the case for the mean response or turnaround time whose calculation will be one of the objectives. Other objectives will be to determine all mean queue lengths, the utilization at each unit and the maximum input rate before saturation.

The equations that describe the total mean arrival rate at each unit are :

$$\gamma_1 = \lambda = 0.7 \qquad \text{IOP}$$

$$\gamma_2 = \gamma_1 + \gamma_3 + 0.9\ \gamma_4 \qquad \text{CPU}$$

$$\gamma_3 = 0.75\ \gamma_2 \qquad \text{Drum}$$

$$\gamma_4 = 0.25\ \gamma_2 \qquad \text{Disk}\ .$$

Solving these equations results in :

$$\gamma_1 = \lambda = 0.7$$

$$\gamma_2 = 40\lambda = 28$$

$$\gamma_3 = 30\lambda = 21$$

$$\gamma_4 = 10\lambda = 7\ .$$

Each service unit can now be analyzed independent of the others. The utilization at each unit is :

$$\rho_1 = \gamma_1/\mu_1 = (0.7)\ (500\ 10^{-3}) = 0.35 \qquad \text{IOP}$$

$$\rho_2 = \gamma_2/\mu_2 = (28)\ (30\ 10^{-3}) = 0.84 \qquad \text{CPU}$$

$$\rho_3 = \gamma_3/\mu_3 = (21)\ (20\ 10^{-3}) = 0.42 \qquad \text{Drum}$$

$$\rho_4 = \gamma_4/\mu_4 = (7)\ (80\ 10^{-3}) = 0.56 \qquad \text{Disk}\ .$$

The mean queue lengths at each unit are :

$$L_1 = \frac{\rho_1}{1 - \rho_1} = 0.54 \qquad \text{IOP}$$

$$L_2 = \frac{\rho_2}{1 - \rho_2} = 5.25 \qquad \text{CPU}$$

$$L_3 = \frac{\rho_3}{1 - \rho_3} = 0.72 \qquad \text{Drum}$$

$$L_4 = \frac{\rho_4}{1 - \rho_4} = 1.27 \qquad \text{Disk} .$$

The mean number of jobs in the network is therefore :

$$L = L_1 + L_2 + L_3 + L_4 = 7.78 .$$

The mean response time can now be calculated by applying Little's Law to the Network, i.e.

$$R = \frac{L}{T} = \frac{7.78}{0.7} = 11.1 \text{ seconds/job} .$$

The maximum input rate can be determined from finding the minimum value of λ that causes one of the unit to reach 100 percent utilization. In terms of the parameter λ the utilizations are :

$$\rho_1 = 0.5 \lambda$$

$$\rho_2 = 1.2 \lambda$$

$$\rho_3 = 0.6 \lambda$$

$$\rho_4 = 0.8 \lambda .$$

Thus, the CPU will saturate first. Setting ρ_2 equal to one and solving for λ results in

$$\lambda_{Max} = 1/1.2 = 0.833 \text{ jobs/s} .$$

If the service discipline at the CPU is changed to processor sharing, then all of the results are the same. Although true this statement is somewhat misleading. The problem is with response time. The means are the same, but the variance distributions are different.

5.4 Arrival Rates Dependent on the Number of Customers in the Network

Six years after his first paper Jackson extended the class of queueing networks that could be solved in a paper that is considered the classic of queueing networks [JACK63]. Again he considered only networks in which the arrival processes from outside the system were Poisson and all service times were exponentially distributed. However, he allowed the mean service time at each server center to vary almost arbitrarily with the number of customer in the service center, and he allowed the mean arrival rate to vary according to the number of customers in the system. In fact the arrival rate could be varied such that if the number of customers fell below some lower limit then a new customer was immediately injected into the system. Also the arrival rate could be set such that if the number of customers in the system reached some upper limit then new customers were not allowed until the number fell below the limit. By setting the upper and lower limits at the same value he considered closed queueing networks. The following is a quote that appeared near the end of the paper 'The discovery of these theorems resulted from making a sequence of guesses concerning more and more general jobshop-like queueing systems, and proving successively more general versions of the theorem'.

It is easier to handle the arrival process if it is assumed that all arrivals emanate from a single Poisson source such that new customers are routed to different service center according to fixed routing probabilities. Recall that if a Poisson source is split then the resulting processes are also Poisson. Therefore, if the arrival process is not a function of the number of customers in the system, then it is equivalent to the multiple Poisson sources in Jackson's first paper. The mean arrival rate at each queue can not be determined before hand if the arrival rate varies according to the number of customers in the system. However, the mean number of visits a customer makes to a service center can be determined. Let e_i represent the mean number of times a customer visits service center i. Then

$$e_i = r_{0i} + \sum_{j=1}^{N} e_j \; r_{ji} \tag{5.22}$$

where r_{0i} is the routing probability that a new customer emanating from the Poisson source visits queue i first.

Jackson proves in his paper that for these more general networks of Markovian queues that the joint probability distribution also satisfies product form. More precisely,

$$P(k_1,\ldots,k_N) = \frac{\lambda(S(K))\; f_1(k_1)\; f_2(k_2)\; \cdots\; f_N(k_N)}{G} \tag{5.23}$$

where

$$f_i(k_i) = \prod_{a=1}^{k_i} e_i/\mu_i(a)$$

$$\lambda(S(K)) = \prod_{a=0}^{K-1} \lambda(a)$$

$$G = \sum_{\substack{\text{all feasible}\\ \text{states}}} \left[\lambda(S(K)) \mathop{}_{i=1}^{N} f_i(k_i)\right]$$

$$K = k_1 + k_2 + \cdots + k_N$$

$\mu_i(a)$ = mean service rate at queue i when it contain 'a' customers

and

$\lambda(a)$ = mean arrival rate when the network contains 'a' customer.

The role of G is that of a normalizing constant to insure that the probabilities sum to one. Of course there is a solution only if G converges to a positive number less than infinity.

As an example of the product form solution consider a two service center network with the following parameters :

$$\lambda(a) = \omega \,/\, (a+1)^x \qquad \text{where } \omega>0 \text{ and } x\geq 0$$

$$\mu_1(a) = b\ a^y \qquad \text{where } b>0 \text{ and } y\geq 0$$

$$\mu_2(a) = c\ a^z \qquad \text{where } c>0 \text{ and } z\geq 0$$

$$r_{01} = r_{12} = r_{20} = 1 \quad \text{the other } r_{ij} = 0 \ .$$

Jackson's theorem states that the joint probability distribution for this network is :

$$P(k_1,k_2) = \frac{(\omega/b)^{k_1}\ (\omega/c)^{k_2}}{G(K)\ ((k_1+k_2)!)^x\ (k_1!)^y\ (k_2!)^z} \ .$$

The following is a proof of Jackson's theorem. By considering all the ways in which state $(k_1,k_2,\ldots,k_N)$ can be reached the set of steady-state equations are :

$$\left[\lambda(K) + \sum_{i=1}^{N} \mu_i(k_i)\right] P(k\ ,\ldots,k\ ) =$$

$$\sum_{i=1}^{N} \lambda(K-1)\ r_{0i}\ P(k_1,\ldots,k_i-1,\ldots k_N)$$

$$+\sum_{i=1}^{N} \mu_i(k_i+1)\ r_{i0}\ P(k_1,\ldots,k_i+1,\ldots,k_N)$$

$$+\sum_{i=1}^{N}\ \sum_{j=1}^{N} \mu_j(k_j+1-\delta_{ij})\ r_{ji}\ P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots k_N). \tag{5.24}$$

Now the following relations are easily seen from the defining equation for $P(k_1,\ldots,k_N)$:

$$\frac{P(k_1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_i,\ldots,k_N)} = \frac{\mu_i(k_i)}{e_i\ \lambda(K-1)} \tag{5.25}$$

$$\frac{P(k_1,\ldots,k_i+1,\ldots,k_N)}{P(k_1,\ldots,k_i,\ldots,k_N)} = \frac{e_i\ \lambda(K)}{\mu_i(k_i+1)} \tag{5.26}$$

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{e_j\ \mu_i(k_i)}{e_i\ \mu_j(k_j+1-\delta_{ij})}\ . \tag{5.27}$$

Dividing both sides of the steady-state equation by $P(k_1,\ldots k_N)$ yields

$$\lambda(K) + \sum_{i=1}^{N} \mu_i(k_i) = \sum_{i=1}^{N} [\mu_i(k_i)\ r_{0i}\ /\ e_i] + \sum_{i=1}^{N} \lambda(K)\ e_i\ r_{i0}$$

$$+\sum_{i=1}^{N}\ \sum_{j=1}^{N} \mu_i(k_i)\ r_{ji}\ e_j\ /\ e_i\ . \tag{5.28}$$

Substituting $r_{0i} = e_i - \sum_{j=1}^{N} e_j\ r_{ji}$ into the first summation on the right and canceling gives :

$$\lambda(K) = \sum_{i=1}^{N} \lambda(K)\ e_i\ r_{i0}\ . \qquad (5.29)$$

Substituting definitions of e_i and r_{i0} shows that the assumed solution balances the steady-state equations.

Several special cases are of sufficient interest to discuss them separately.

5.4.1 The Constant Arrival Rate Case

If the arrival rate does not depend on the number of customers, then for all 'a', $\lambda(a) = \lambda$ = constant and $\lambda(S(K)) = \lambda^{K-1}$. For this case the steady-state probabilities are

$$P(k_1,\ldots,k_N) = \frac{\lambda^{K-1}\ f_1(k_1)\ f_2(k_2)\ \ldots\ f_N(k_N)}{G} \qquad (5.30)$$

where $f_i(k_i)$, K, and G are the same as before. Multiplying the numerator and denominator by λ and letting G absorb the λ in the denominator results in

$$P(k_1,\ldots,k_N) = \frac{[\lambda^{k_1}\ f_1(k_1)]\ [\lambda^{k_2}\ f_2(k_2)]\ \ldots\ [\lambda^{k_N}\ f_N(k_N)]}{G}\ . \qquad (5.31)$$

Let $g_i(k_i) = \lambda^{k_i}\ f_i(k_i)$. That is,

$$g_i(k_i) = \prod_{a=0}^{k_i} \lambda\ e_i/\mu_i(a)\ . \qquad (5.32)$$

In terms of $g_i(k_i)$ the steady-state probabilities are

$$P(k_i,\ldots,k_N) = \frac{g_1(k_1)\ g_2(k_2)\ \ldots\ g_N(k_N)}{G}\ . \qquad (5.33)$$

Since the network is open all states are feasible and

$$G = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \cdots \sum_{k_N=0}^{\infty} \left[g_1(k_1)\, g_2(k_2) \,\ldots\, g_N(k_N) \right]$$

$$= \left[\sum_{k_1=0}^{\infty} g_1(k_1) \right] \left[\sum_{k_2=0}^{\infty} g_2(k_2) \right] \cdots \left[\sum_{k_N=0}^{\infty} g_N(k_N) \right] . \qquad (5.34)$$

Notice that G factors into terms where each term involves parameters for a single service center. It follows that $P(k_1,\ldots,k_N)$ factors into terms that involve a single service center. That is

$$P(k_1,\ldots,k_N) = P_1(k_1)\; P_2(k_2)\; \ldots\; P_N(k_N)$$

where

$$P_i(k_i) = g_i(k_i) \;/ \sum_{k_i=0}^{\infty} g_i(k_i) \quad . \qquad (5.35)$$

Thus, the distribution of customers at the service centers are independent, and the distribution at each center is the same as for a one-service-center queueing system where customers arrive from a Poisson process with mean rate $\lambda\, e_i$, and the service completion process is identical to that of service center i. Also, since e_i is the mean number of times a customer visits service center i, it follows that for this case $\lambda\, e_i$ is indeed the mean arrival rate of customers to center i. Furthermore if center i contains m_i servers and

$$\mu_i(k_i) = \begin{cases} k_i\, \mu_i & \text{for } k_i \leq m_i \\ m_i\, \mu_i & \text{for } k_i \geq m_i \,, \end{cases} \qquad (5.36)$$

then the results are of course the same as in section 5.2.

5.4.2 Closed Networks

A closed network is one in which the number of customers in the network remains constant. Jackson considered closed networks as a special case of networks in which the arrival rate varies according to the number of customers in the network. To keep the number of customers at some constant value K, he set $\lambda(k) = 0$ for $k \geq K$ and $\lambda(K-1)=\infty$. Thus if the number of customers falls below K a new customer is immediately injected into the network, and if the number of customers is K, new customers cannot enter the network. This is equivalent to a network in which the same customers circulate eternally.

Unaware of Jackson's work Gorden and Newell published a paper on closed networks. It appeared four years later in the same periodical in which Jackson's paper appeared [GORD67]. They acknowledged later that their formulae for steady-state probabilities could be obtained by specializing the parameters of Jackson's more general model. However, researchers at the time were unaware of this and treated Gordon and Newell's simplified notation and results as independent work. Even today credit is more often given to Gordon and Newell.

The approach used here is similar to that used by Jackson. For a closed network $\lambda(S(K))$ is assigned the value of one, and since r_{i0} equals zero for all i,

$$e_i = \sum_{j=1}^{N} e_j \, r_{ji} \quad . \tag{5.37}$$

By considering all the ways in which state $(k_1,k_2,\ldots,k_N)$ can be reached the steady-state equation is

$$\sum_{i=1}^{N} \mu_i(k_i)\ P(k_1,\ldots,k_N) =$$

$$\sum_{i=1}^{N}\ \sum_{j=1}^{N} \mu_j(k_j+1-\delta_{ij})\ r_{ji}\ P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)\cdot \tag{5.38}$$

From the defining equation

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{e_j\ \mu_i(k_i)}{e_i\ \mu_j(k_j+1-\delta_{ij})}\ . \tag{5.39}$$

Dividing both sides by $P(k_1,\ldots,k_N)$ results in

$$\sum_{i=1}^{N} \mu_i(k_i) = \sum_{i=1}^{N}\ \sum_{j=1}^{N} \mu_i(k_i)\ r_{ji}\ e_j\ /\ e_i\ \ . \tag{5.40}$$

Using the definition of e_i or more precisely the fact that

$$\sum_{j=1}^{N} [r_{ji}\ e_j\ /\ e_i] = 1 \tag{5.41}$$

shows the equation is balanced. Hence the solution satisfies the steady-state equations.

Again not only are the steady-state equations balanced, but it turns out that

$$\mu_i(k_i)\ P(k_1,\ldots,k_N) =$$

$$\sum_{j=1}^{N} \mu_j(k_j+1-\delta_{ij})\ r_{ji}\ P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)\ . \tag{5.42}$$

The implication is that local balance also applies to closed networks.

Notice that if all of the e's are multiplied by a constant, then the set of N equations defining them is still satisfied. Hence, for a closed network there are only N-1 independent equations. The solution is to assign one of the e's an arbitrary positive value. Only their ratios appeared in the proof. The e's in a closed network are often referred to as relative throughputs.

Although from a theoretical point of view any positive value can be assigned to one of the e's. The value selected does affect the normalizing constant, G, and can cause numerical problems such as overflow or underflow. Compensating for the fact that the magnitudes are not know is only one of the purposes of the normalizing constant. It would still be required even if the magnitudes were known before hand. The normalizing constant, G, is virtually a function of every parameter in the network.

It is remarkable that joint probability distribution of a closed network of queues with exponential servers has a product form solution. That is, the form of the solution is the product of N queues with Poisson arrivals and exponential servers, divided by a normalizing constant. What makes this so remarkable is that none of the arrival processes are Poisson at any service center. Again no one has an explanation of why this is so.

Even though the solution has a product form the distributions at the individual service centers are not independent since their sum must always equal the same value. This is the primary reason that a closed form solution for the normalizing constant can not be obtained.

Determining the normalizing constant by the obvious way of finding all of the unnormalized probabilities can be and usually is a difficult problem. By considering the number of possible ways that customers can be distributed in a closed network the number of states can be determined. This problem is equivalent to that of finding the number of permutations of N-1+K objects of which N-1 are the same and K are the same. Thus for a closed network with N service centers and K customers the number of states is

$$\frac{(N + K - 1)!}{(N - 1)!\ K!} = \binom{N + S(K) - 1}{N-1}. \qquad (5.43)$$

For example for a network with 8 eight service centers and 20 customer there are 888,030 states. Fortunately other techniques to determine the normalizing constant exist. One of these will be discussed in detail in Chapter 7.

5.4.3 An Application of Closed Queueing Networks

As an example of a closed network consider the model in Figure 5.7. Again assume that each service center has a single server, and that all jobs are statistically identical, and that all service times are exponentially distributed. The parameters for the network are given in table 5.2. A job making a CPU to CPU transition is regarded as having left the system and having been immediately replaced by another job. Thus, the flow along the CPU to CPU path represents the system throughput. The objective is to determine the mean number of customers at each service center, the utilization of each service center, the mean throughput, and mean response or turnaround time of a job. To make

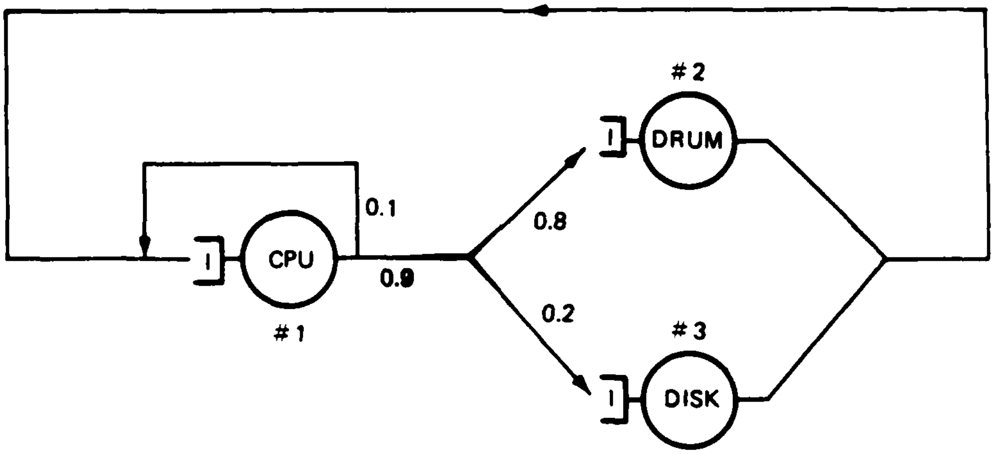


Figure 5.7 Closed Network Central Server Model.

Parameter name	Symbol	Value
Mean uninterrupted CPU time	$1/\mu_1$	10ms
Mean drum service time	$1/\mu_2$	25ms
Mean disk service time	$1/\mu_3$	100ms
CPU to CPU probability	r_{11}	0.1
CPU to drum probability	r_{12}	0.8
CPU to disk probability	r_{13}	0.1
Drum to CPU probability	r_{21}	1.0
Disk to CPU probability	r_{31}	1.0

Table 5.2 Parameters for Figure 5.7.

the problem manageable assume that the number of jobs in the system is three.

Since there are three service centers and three jobs in the network the total number of network states is

$$\binom{3+3-1}{3-1} = \frac{5!}{2!\ 3!} = 10 \ .$$

The state-transition-rate diagram for this network is depicted in Figure 5.8. The ten steady-state or global balance equations are:

$$\mu_1 \ P(3,0,0) = \mu_1 \ r_{11} \ P(3,0,0) + \mu_2 \ P(2,1,0) + \mu_3 \ P(2,0,1)$$

$$\begin{aligned}[\mu_1+\mu_3] \ P(2,0,1) = {} & \mu_1 \ r_{11} \ P(2,0,1) + \mu_1 \ r_{13} \ P(3,0,0) \\ & + \mu_2 \ P(1,1,1) + \mu_3 \ P(1,0,2)\end{aligned}$$

$$\begin{aligned}[\mu_1+\mu_2] \ P(2,1,0) = {} & \mu_1 \ r_{11} \ P(2,1,0) + \mu_1 \ r_{12} \ P(3,0,0) \\ & + \mu_2 \ P(1,1,1) + \mu_3 \ P(1,2,0)\end{aligned}$$

$$\begin{aligned}[\mu_1+\mu_2] \ P(1,0,2) = {} & \mu_1 \ r_{11} \ P(1,0,2) + \mu_1 \ r_{13} \ P(2,0,1) \\ & + \mu_2 \ P(0,1,2) + \mu_3 \ P(0,0,3)\end{aligned}$$

$$\begin{aligned}[\mu_1+\mu_2+\mu_3] \ P(1,1,1) = {} & \mu_1 \ r_{11} \ P(1,1,1) + \mu_1 \ r_{12} \ P(2,0,1) \\ & + \mu_1 \ r_{13} \ P(2,1,0) + \mu_2 \ P(0,2,1) + \mu_3 \ P(0,1,2)\end{aligned}$$

$$\begin{aligned}[\mu_1+\mu_2] \ P(1,2,0) = {} & \mu_1 \ r_{11} \ P(1,2,0) + \mu_1 \ r_{12} \ P(2,1,0) \\ & + \mu_2 \ P(0,3,0) + \mu_3 \ P(0,2,1)\end{aligned}$$

$$\mu_3 \ P(0,0,3) = \mu_1 \ r_{13} \ P(1,0,2)$$

$$[\mu_2+\mu_3] \ P(0,1,2) = \mu_1 \ r_{12} \ P(1,0,2) + \mu_1 \ r_{13} \ P(1,1,1)$$

$$[\mu_2+\mu_3] \ P(0,2,1) = \mu_1 r_{12} \ P(1,1,1) + \mu_1 r_{13} \ P(1,2,0)$$

$$\mu_2 \ P(0,3,0) = \mu_1 r_{12} \ P(1,2,0) \ .$$

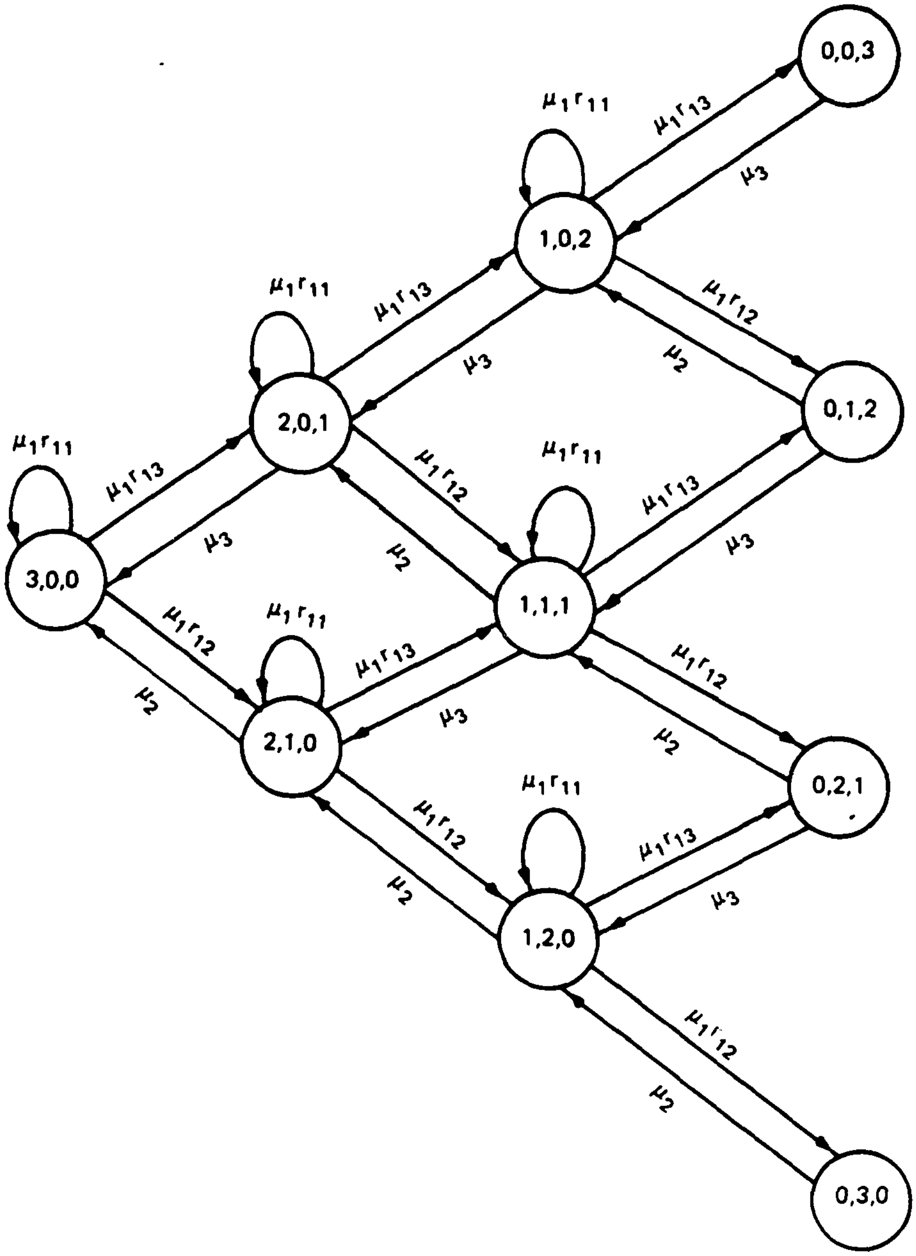


Figure 5.8 State-Transition-Rate Diagram for the Central Server Model with Three Jobs.

Of course the steady-state probabilities could be determined by solving these ten simultaneous equations. Another approach would be to write local balance equations and solve these. For example the local balance equations corresponding to the fifth global balance equation are:

$$\mu_1 \ P(1,1,1) = \mu_1 \ r_{11} \ P(1,1,1) + \mu_2 \ P(0,2,1) + \mu_3 \ P(0,1,2)$$
$$\mu_2 \ P(1,1,1) = \mu_1 \ r_{12} \ P(2,0,1)$$
$$\mu_3 \ P(1,1,1) = \mu_1 \ r_{13} \ P(2,1,0) \ .$$

Neither of these approaches will be used here. The method described by Jackson will be used instead.

The equations describing the mean number of times a job visits a service center are :

$$e_1 = 0.1 \ e_1 + e_2 + e_3$$
$$e_2 = 0.8 \ e_1$$
$$e_3 = 0.1 \ e_1 \ .$$

Observe that there are only two independence equations. Although any positive value can be assigned to one of the e's, assigning 100 to e_1 causes the ratio of e_i/μ_i to be integers. Selerting e_1 as 100 results in :

$$e_1 = 100 \quad \text{and} \quad e_1/\mu_1 = 1$$
$$e_2 = 80 \quad \text{and} \quad e_2/\mu_2 = 2$$
$$e_3 = 10 \quad \text{and} \quad e_3/\mu_3 = 1 \quad .$$

The joint probability distribution that service center one contains k_1 jobs, service center two k_2 jobs, and service center three k_3 jobs is

$$P(k_1,k_2,k_3) = \frac{f_1(k_1)\ f_2(k_2)\ f_3(k_3)}{G}$$

where $$f_i(k_i) = \prod_{a=1}^{k_i} (e_i/\mu_i) = (e_i/\mu_i)^{k_i} .$$

Hence, $$P(k_1,k_2,k_3) = \frac{(e_1/\mu_1)^{k_1}\ (e_2/\mu_2)^{k_2}\ (e_3/\mu_3)^{k_3}}{G} .$$

The ten steady—state probabilities are therefore:

$$P(1,1,1) = 2/G$$

$$P(1,0,2) = 1/G$$

$$P(2,0,1) = 1/G$$

$$P(0,1,2) = 2/G$$

$$P(0.0,3) = 1/G$$

$$P(0,2,1) = 4/G$$

$$P(2,1,0) = 1/G$$

$$P(3,0,0) = 1/G$$

$$P(1,2,0) = 4/G$$

$$P(0,3,0) = 8/G .$$

The normalizing constant is determine from the fact the the probabilities must sum to one. More precisely,

$$\sum_{\text{all states}} P(k_1,k_2,k_3) = 26/G = 1 .$$

Hence, G = 26. The mean queue length at each service center is

$$L_i = \sum_{\text{all states}} k_i \, P(k_1,k_2,k_3) \; .$$

$$L_1 = (1)\,P(1,1,1) + (1)\,P(1,0,2) + (2)\,P(2,0,1) + (2)\,P(2,1,0)$$

$$+ (3)\,P(3,0,0) + (1)\,P(1,2,0)$$

$$= [(1)(2) + (1)(1) + (2)(1) + (2)(2) + (3)(1) + (1)(4)] \;/\; 26$$

$$= 0.615$$

$$L_2 = (1)\,P(1,1,1) + (1)\,P(0,1,2) + (2)\,P(0,2,1) + (1)\,P(2,1,0)$$

$$+ (2)\,P(1,2,0) + (3)\,P(0,3,0)$$

$$= (1)(2) + (1)(2) + (2)(4) + (1)(2) + (2)(4) + (3)(8)] \;/\; 26$$

$$= 1.769$$

$$L_3 = (1)\,P(1,1,1) + (2)\,P(1,0,2) + (1)\,P(2,0,1) + (2)\,P(0,1,2)$$

$$+ (3)\,P(0,0,3) + (1)\,P(0,2,1)$$

$$= [(1)(2) + (2)(1) + (1)(1) + (2)(2) + (3)(1) + (1)(4)] \;/\; 26$$

$$= 0.615$$

Notice that $L_1 + L_2 + L_3 = 3$ as expected.

The utilization of a service center equals the probability that the service center contains at least one customer, which is one minus the probability that the service center contains zero customers. The marginal probability that the service centers contain zero customers is:

$$P_1(0) = P(0,1,2) + P(0,0,3) + P(0,2,1) + P(0,3,0)$$

$$= (2 + 1 + 4 + 8) \;/\; 26 \;\; = 0.577$$

$$P_2(0) = P(1,0,2) + P(2,0,1) + P(0,0,3) + P(3,0,0)$$

$$= (1 + 1 + 1 + 1) \;/\; 26 \;\; = 0.154$$

$$P_3(0) = P(2,1,0) + P(3,0,0) + P(1,2,0) + P(0,3,0)$$
$$= (2 + 1 + 4 + 8) \;/\; 26 \;\; = 0.577 \;.$$

The utilization at the three service centers is :

$$\rho_1 = 1 - P_1(0) = 0.423$$
$$\rho_2 = 1 - P_2(0) = 0.846$$
$$\rho_3 = 1 - P_3(0) = 0.423 \;.$$

The throughput of the system can be determined from the utilization and mean service time at service center one. If the service center one was busy 100 percent of the time the number of customer served per second would be 1/10ms, or 100 jobs per second. Utilization equals the long run percent that the service center is busy. Hence the number of jobs passing through service center one is 0.423 x 100 = 42.3 jobs/s. The probability that a job makes a CPU to CPU transition is 0.1, therefore the number of customer completing service (the throughput) is 4.23 jobs/s. The mean response time can be determine by applying Little's Law to the network. That is,

$$R = \frac{L}{T} = \frac{3}{4.23} = 0.798 \text{ seconds per job.}$$

5.5.4 <u>Open Networks with Finite Storage Capacity</u>

Although the form of the solution for a system with arrival rates dependent on the number of customers in the network was given in section 5.5, the normalizing constant is next to impossible to determine unless there exists a positive integer K^* such that $\lambda(K) = \lambda(K^*)$ for $K \geq K^*$. The most interesting case is when $\lambda(K^*) = 0$. This is because almost all real systems can contain only a finite number of customer. Customers that arrive when the system is full are simply turned away without receiving service. A common example of this is the telephone system. In fact it was the study of this system that brought about the birth of queueing theory.

A queueing network with finite storage capacity is equivalent to a closed network with K^* customers. The sources and sinks in the original network are replaced by a service center with rate $\mu(k)=\lambda(K^*-k)$. Since the interdeparture process (time between departures) of this service center is the same as the interarrival process (time between arrivals) of the original network, the two networks are equivalent. This is best illustrated by example. The network in Figure 5.9 is equivalent to a single service center with Poisson arrival rate μ_1 and finite storage capacity K^*. To see this first recall that the steady-state solution is invariant to the initial distribution of customers. Now assume that all K^* customers are initially at the first service center. As long as there are customers at this service center the departure process is

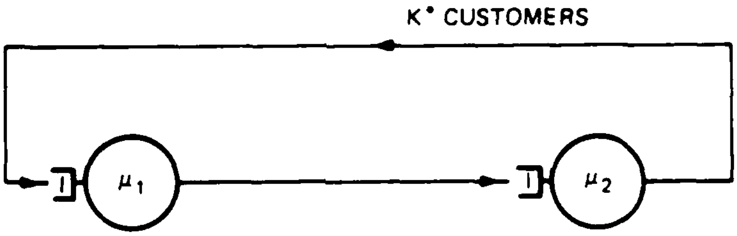


Figure 5.9 Equivalent $M/M/1/K^*$ System.

exponentially distributed with mean μ_1. Now since the departure process at the first service center is the arrival process at the second service center, it follows that the arrival process is Poisson with mean rate,

$$\lambda(k_2) = \begin{cases} \mu_1 & \text{for } k_2 < K^* \\ 0 & \text{for } k_2 = K^* \end{cases} .$$

Which is identical to

$$\mu_1(k_1) = \lambda(K^* - k_1) = \begin{cases} \mu_1 & \text{for } k_1 > 0 \\ 0 & \text{for } k_1 = 0 \end{cases} .$$

Hence the first service center controls the arrival process to the second service center, and the network is equivalent to the $M/M/1/K^*$ - system. Note that $K = k_2$ in this example.

A more mathematical proof that the two systems are the equivalent consists of showing that the probability distribution of customers is the same. More precisely,

$$P(k_2) = P(k_1, k_2) = \frac{\left[\frac{1}{\mu_1}\right]^{k_1} \left[\frac{1}{\mu_2}\right]^{k_2}}{G} = \frac{\left[\frac{1}{\mu_1}\right]^{K^*} \left[\frac{\mu_1}{\mu_2}\right]^{k_2}}{G} .$$

Solving for the normalizing constant yields:

$$G = \sum_{k_2=0}^{K^*} (1/\mu_1)^{K^*} (\mu_1/\mu_2)^{k_2}$$

$$= \left[\frac{1}{\mu_1}\right]^{K^*} \frac{[1-(\mu_1/\mu_2)]^{K^*+1}}{[1-(\mu_1/\mu_2)]} .$$

Thus,

$$P(k_2) = \frac{(\mu_1/\mu_2)^{k_2}\ [1-(\mu_1/\mu_2)]}{[1-(\mu_1/\mu_2)]^{K^*+1}} \quad ,$$

which is identical of that for the M/M/1/K* system with $\lambda=\mu_1$ and $k=k_2$.

An example where the customer population is finite and the arrival rate varies according to the number of customers at the system is the M/M/1//M system (Chapter 3, Section 3.8). The system is equivalent to the network in Figure 5.10. Recall that for this system the arrival rate for the M/M/1//M system is

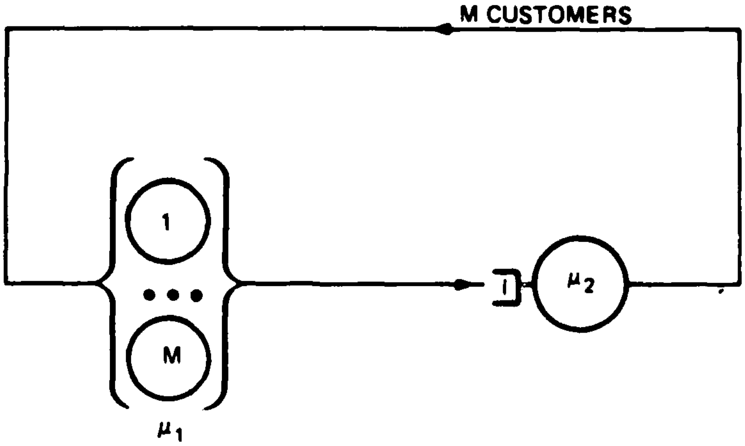


Figure 5.10 Equivalent M/M/1//M System.

$$\lambda_k = \begin{cases} (M-k)\lambda & \text{for } k<M \\ 0 & \text{for } k=M \ . \end{cases}$$

Thus, $K^*=M$ and $K=k=k_2$ and hence

$$\lambda(k_2) = \begin{cases} (M-k_2)\lambda & \text{for } k_2<M \\ 0 & \text{for } k_2=M \ , \end{cases}$$

or equivalently

$$\bar{\mu}_1(k_1)=\lambda(M-k_1) = \begin{cases} k_1\ \mu_1 & \text{for } k_1>0 \\ 0 & \text{for } k_1=0 . \end{cases}$$

Note that in order to show that the probability distributions are the same M was not replaced by K^*. It follow that:

$$P(k_2) = P(k_1,k_2) = \frac{\frac{1}{k_1!}\left[\frac{1}{\mu_1}\right]^{k_1} \left[\frac{1}{\mu_2}\right]^{k_2}}{G} = \frac{\frac{1}{(M-k_2)!}\left[\frac{1}{\mu_1}\right]^{M} \left[\frac{\mu_1}{\mu_2}\right]^{k_2}}{G} ,$$

and

$$G = \sum_{k_2=0}^{M} [1/(M-k_2)!]\ (1/\mu_1)^M\ (\mu_1/\mu_2)^{k_2}$$

$$= (1/\mu_1)^M \sum_{k_2=0}^{M} [1/(M-k_2)!]\ (\mu_1/\mu_2)^{k_2} ,$$

and finally

$$P(k_2) = \frac{\frac{M!}{(M-k_2)!}\ \left[\frac{\mu_1}{\mu_2}\right]^{k_2}}{\sum_{k_2=0}^{M} [M!/(M-k_2)!]\ (\mu_1/\mu_2)^{k_2}} ,$$

which is the same as the M/M/1//M system with $\lambda=\mu_1$ and $k=k_2$.

In general any network with finite storage capacity can be mapped by this procedure into an equivalent closed network.

5.3.5 Service Rates and Subsets of Service Centers

For a subset of the service centers the service rate may be a function of both the number of customers in the service center and the number in the subset [BASK75]. The following assumes the service centers are numbered such that 1,2,...,M are in the subset. Let K_I be the number of customers in the subset and let $Z_I(K_I)$ be a positive function such that the service rate of the centers in the subset is $\mu_i(k_i)\ Z_I(K_I)$. For this case the joint distribution is the same as before except

$$\prod_{i=1}^{M} f_i(k_i) \;:=\; \prod_{i=1}^{M} f_i(k_i) \prod_{a=1}^{K_I} (1/Z_I(a)) \quad , \qquad (5.44)$$

where the symbol := is an assignment operator and is read as 'becomes'.

Since this will be used later, and since it is not known to be proved elsewhere, it will be proved here. By considering all the ways in which state $(k_1,k_2,\ldots,k_N)$ may be reached the steady-state equation is:

$$\left[\lambda(K) + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) + \sum_{i=M+1}^{N} \mu_i(k_i)\right] P(k_1,\ldots,k_N) =$$

$$\sum_{i=1}^{M} \lambda(K-1)\ r_{0i}\ P(k_1,\ldots,k_i-1,\ldots,k_N)$$

$$+ \sum_{i=M+1}^{N} \lambda(K-1)\ r_{0i}\ P(k_1,\ldots,k_i-1,\ldots,k_N)$$

$$
\begin{aligned}
&+ \sum_{i=1}^{M} \mu_i(k_i+1)\; Z_I(k_I+1)\; r_{i0}\; P(k_1,\dots,k_i+1,\dots,k_N) \\
&+ \sum_{i=M+1}^{N} \mu_i(k_i+1)\; r_{i0}\; P(k_1,\dots,k_i+1,\dots,k_N) \\
&+ \sum_{i=1}^{M} \sum_{j=1}^{M} \mu_j(k_j+1-\delta_{ij})\; Z_I(k_I)\; r_{ji}\; P(k_1,\dots,k_j+1,\dots,k_i-1,\dots,k_N) \\
&+ \sum_{i=M+1}^{N} \sum_{j=1}^{M} \mu_j(k_j+1)\; Z_I(k_I+1)\; r_{ji}\; P(k_1,\dots,k_j+1,\dots,k_i-1,\dots,k_N) \\
&+ \sum_{i=1}^{M} \sum_{j=M+1}^{N} \mu_j(k_j+1)\; r_{ji}\; P(k_1,\dots,k_j+1,\dots,k_i-1,\dots,k_N) \\
&+ \sum_{i=M+1}^{N} \sum_{j=M+1}^{N} \mu_j(k_j+1-\delta_{ij})\; r_{ji}\; P(k_1,\dots,k_j+1,\dots,k_i-1,\dots,k_N)\;.
\end{aligned}
\tag{5.45}
$$

The following relations are from the defining equation of $P(k_1,\dots,k_N)$:

$$\frac{P(k_1,\dots,k_i-1,\dots,k_N)}{P(k_1,\dots,k_i,\dots,k_N)} = \frac{\mu_i(k_i)\; Z_I(K_I)}{e_i\; \lambda(K-1))} \qquad \text{for } i \leq M$$

$$\frac{P(k_1,\dots,k_i-1,\dots,k_N)}{P(k_1,\dots,k_i,\dots,k_N)} = \frac{\mu_i(k_i)}{e_i\; \lambda(K-1))} \qquad \text{for } i > M$$

$$\frac{P(k_1,\dots,k_i+1,\dots,k_N)}{P(k_1,\dots,k_i,\dots,k_N)} = \frac{e_i\; \lambda(K)}{\mu_i(k_i+1)\; Z_I(k_I+1)} \qquad \text{for } i \leq M$$

$$\frac{P(k_1,\ldots,k_i+1,\ldots,k_N)}{P(k_1,\ldots,k_i,\ldots,k_N)} = \frac{e_i\ \lambda(K)}{\mu_i(k_i+1)} \qquad \text{for } i>M$$

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{e_j\ \mu_i(k_i)}{e_i\ \mu_j(k_j+1-\delta_{ij})} \qquad \begin{array}{ll}\text{for} & i,j>M \\ \text{or} & i,j<M\end{array}$$

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{e_j\ \mu_i(k_i)}{e_i\ \mu_j(k_j+1)\ Z_I(k_I+1)} \qquad \begin{array}{ll}\text{for} & j<M \\ \text{and} & i\geq M\end{array}$$

$$\frac{P(k_1,\ldots,k_j+1,\ldots,k_i-1,\ldots,k_N)}{P(k_1,\ldots,k_j,\ldots,k_i,\ldots,k_N)} = \frac{e_j\ \mu_i(k_i)\ Z_I(K_I)}{e_i\ \mu_j(k_j+1)} \qquad \begin{array}{ll}\text{for} & j>M \\ \text{and} & i\leq M\ .\end{array}$$

(5.46)

Following the usual procedure of dividing both sides by $P(k_1,\ldots,k_N)$ and using these relations results in:

$$\lambda(K) + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) + \sum_{i=M+1}^{N} \mu_i(k_i) =$$

$$+ \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I)\ r_{0i}\ /\ e_i \ + \sum_{i=M+1}^{N} \mu_i(k_i)\ r_{0i}\ /\ e_i$$

$$+ \sum_{i=1}^{M} \lambda(K)\ r_{i0}\ e_i \ + \sum_{i=M+1}^{N} \lambda(K)\ r_{i0}\ e_i$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{M} \mu_i(k_i)\ Z_I(k_I)\ r_{ji}\ e_j\ /\ e_i \ + \sum_{i=M+1}^{N} \sum_{j=1}^{M} \mu_i(k_i)\ r_{ji}\ e_j\ /\ e_i$$

$$+ \sum_{i=1}^{M} \sum_{j=M+1}^{N} \mu_i(k_i)\ Z_I(k_I)\ r_{ji}\ e_j\ /\ e_i \ + \sum_{i=M+1}^{N} \sum_{j=M+1}^{N} \mu_i(k_i)\ r_{ji}\ e_j\ /\ e_i$$

. (5.47)

Combining the third and fourth, fifth and seventh, sixth and eighth, terms on the right yields:

$$\lambda(K) + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) + \sum_{i=M+1}^{N} \mu_i(k_i) =$$

$$+ \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I)\ r_{0i} / e_i + \sum_{i=M+1}^{N} \mu_i(k_i)\ r_{0i} / e_i$$

$$+ \sum_{i=1}^{N} \lambda(K)\ r_{i0}\ e_i$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{N} \mu_i(k_i)\ Z_I(k_I)\ r_{ji}\ e_j / e_i + \sum_{i=M+1}^{N} \sum_{j=1}^{N} \mu_i(k_i)\ r_{ji}\ e_j / e_i\ . \tag{5.48}$$

Substituting $r_{0i} = e_i - \sum_{j=1}^{N} e_j\ r_{ji}$ yields :

$$\lambda(K) + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) + \sum_{i=M+1}^{N} \mu_i(k_i) =$$

$$+ \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) - \sum_{i=1}^{M} \sum_{j=1}^{N} \mu_i Z_I(k_I)\ r_{ji}\ e_j / e_i$$

$$+ \sum_{i=M+1}^{N} \mu_i(k_i) - \sum_{i=M+1}^{M} \sum_{j=1}^{N} \mu_i(k_i)\ r_{ji}\ e_j / e_i$$

$$+ \sum_{i=1}^{N} \lambda(K)\ r_{i0}\ e_i$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{N} \mu_i(k_i)\ Z_I(k_I)\ r_{ji}\ e_j\ /\ e_i\ + \sum_{i=M+1}^{N} \sum_{j=1}^{N} \mu_i(k_i)\ r_{ji}\ e_j\ /\ e_i \tag{5.49}$$

or equivalently,

$$\lambda(K) + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I) + \sum_{i=M+1}^{N} \mu_i(k_i) =$$

$$+ \sum_{i=1}^{N} \lambda(K)\ r_{i0}\ e_i\ + \sum_{i=1}^{M} \mu_i(k_i)\ Z_I(k_I)\ + \sum_{i=M+1}^{N} \mu_i(k_i)\ . \tag{5.50}$$

Finally substituting

$$r_{i0} = \sum_{j=1}^{N} r_{ij} \quad \text{and} \quad e_i = r_{0i} + \sum_{j=1}^{N} e_j\ r_{ji}$$

balances the equation.

In the next chapter more advanced networks will be covered.

CHAPTER 6

ADVANCED QUEUEING NETWORKS

6.1 Customer Classes

In the queueing networks discussed in the last chapter, it was assumed that all customers at a service center were identical. The usual way to eliminate this assumption is to partition the customers at a service center into classes. Within a class all customers are homogeneous, but different classes may have different service time distributions, priorities, routing, etc. It is important to emphasize that classes are associated with service centers, and that customers are distinguished at the service center level. This is more general and includes the case in which customers are distinguished at the network level. For example, all customers entering a network may be identical, whereas a customer visiting a service center for the second time may have a different service time distribution and routing probabilities than a customer that is visiting the same service center for the first time. Thus, at the service center level the two customers behave differently.

The notation used in the last chapter is easily extended to include classes. A customer at service center i in class s, after receiving service, proceeds to service center j class t according to the routing probability $r_{is:jt}$. The mean service time of a customer at service center j in class t is denoted μ_{jt}.

Figure 6.1 shows a closed network with two service centers and three classes. Service center two is represented by a service facility with two queues, one for each class. This depicts the situation where service time distributions and routing may be different for different classes. In reality there is only one queue at service center two.

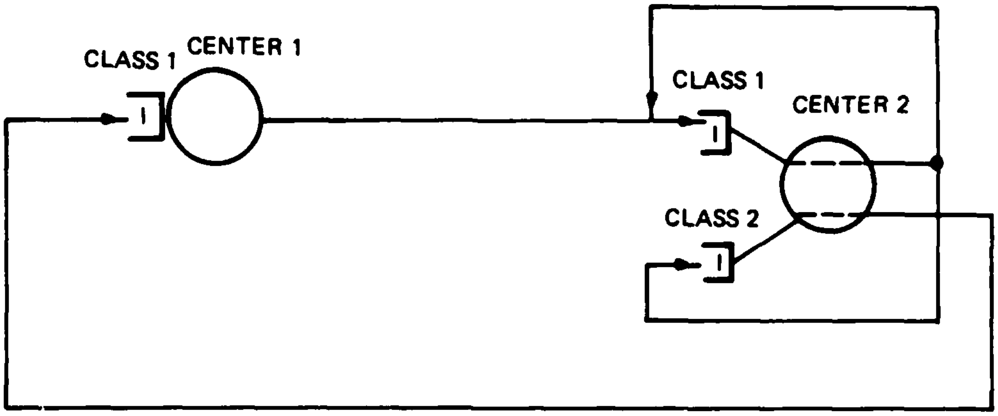


Figure 6.1 Network with Two Service Centers and Three Classes.

Figure 6.2 shows a closed network with two routing chains. The routing probabilities are such that a customer in the top chain (loop) cannot make a transition to a class in the bottom chain. The same is true of customers in the bottom chain. Hence, the number of customers in the top and bottom chains are both constants. Usually, customers in different routing chains are distinguishable at the network level.

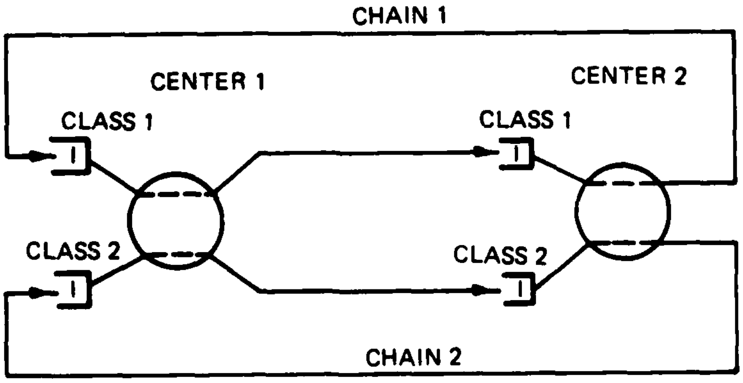


Figure 6.2 Closed Network with Two Chains.

Figure 6.3 depicts an open network with two routing chains. The number of customers in each chain is a random variable. It is assumed that both sources are Poisson. However the mean arrival rate of customers from source one may depend on the number of customers in chain one (the top chain). The same applies to the mean arrival rate of customers from source two. If the mean arrival rates are constant, which is the usual case, then the sources can be combined into a single Poisson source, and the network reduces to a single chain. In formulating open queueing network models, it is often convenient to assume that there are multiple routing chains. However, for computational purposes, it is desirable to have only one chain. It is also true that if classes are used only for routing purposes, then generally, the number of classes required can be reduced by combining the sources.

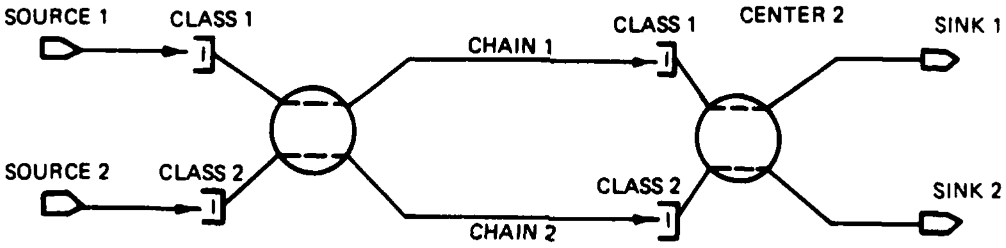


Figure 6.3 An Open Network with Two Chains.

Figure 6.4 depicts a mixed network. The network is a combination of an open and closed network. The number of customers in the top chain is a random variable, whereas the number of customers in the bottom chain is constant.

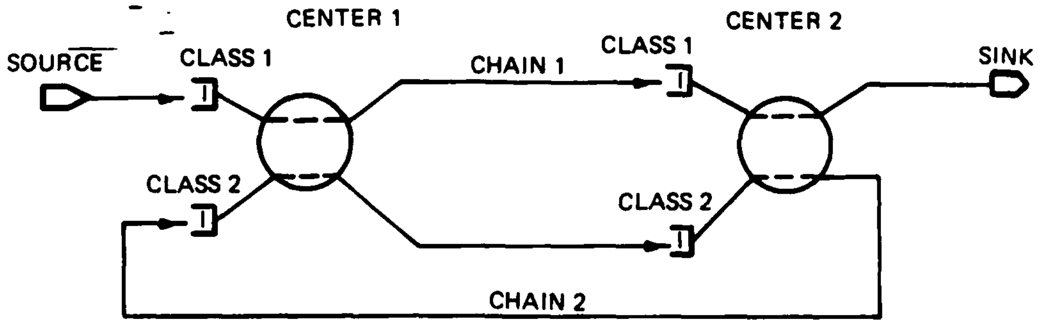


Figure 6.4 Mixed Network with Two Chains.

6.2 Nonexponential Service Times

Another limitation of the queueing networks in the preceding chapter was that all service times had to be exponentially distributed. The only approach for dealing with nonexponential service times is to represent them by a combination of series and parallel stages in which the time spent in each stage is an independent random variable that is exponentially distributed. The necessary condition to accomplish this is usually stated as : the probability distribution function must have a rational Laplace transform (can be expressed as the ratio of two polynomials in s) [BASK75] [KLEI76] [KOBA81] [HAYE84]. This statement is conditionally correct, however, in order for the probability distribution function to have a rational Laplace transform it is necessary that the probability density function have one also. More precisely,

$$b^*(s) = sB^*(s) - B(0^-) \quad , \qquad (6.1)$$

where

$b^*(s)$ = Laplace transform of the density function,

$B^*(s)$ = Laplace transform of the distribution function,

and

$B(0^-)$ = The distribution function evaluated at 0^-.

The reason for attaching such significance to this point, is that there are no examples in the literature, and the statement leads one to believe that it is the distribution function that is expanded to obtain the stages, when in fact it is the density function. The crucial point is that the joint density function of the sum of two independent random variables is the convolution of the individual density functions, and therefore the Laplace transform of the joint density function is the product of the individual transforms.

The procedure then consists of taking the Laplace transform of the density function and expanding it into a series of exponential stages. The obvious way to perform the expansion is to use the method of ordinary partial fraction expansion. However, this does not yield the minimum number of stages, and the computation complexity rises rapidly with each stage. Although it may not be well known, there are several versions of partial fractions [COX55]. Any density function that has a rational Laplace transform can be expanded in the following form :

$$b^*(s) = b_0 + \sum_{j=1}^{z} a_0 \ldots a_{j-1} b_j \prod_{i=1}^{j} \mu_i/(s+\mu_i) \qquad (6.2)$$

where z is the order of the denominator, and $a_i+b_i=1$, for $i=0,1,\ldots,z-1$ and $b_z=1$. The structure of the representation that results from this type expansion is depicted in Figure 6.5. The number of stages always equals the order of the denominator. Therefore, when there are repeated roots, it results in fewer exponential stages than would ordinary partial fractions .

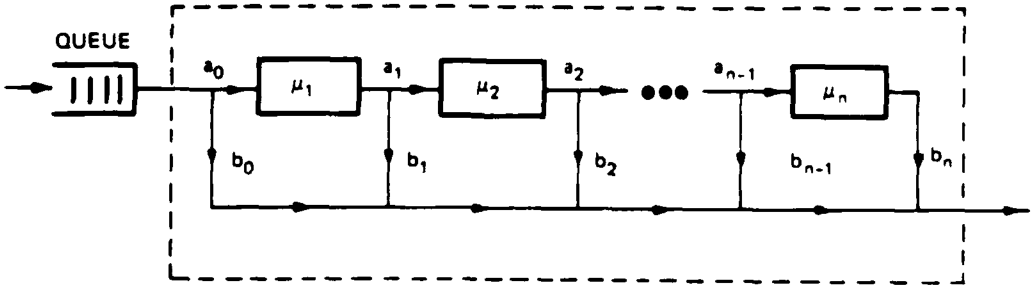


Figure 6.5 Cox's Method of Exponential Stages.

When a customer arrives at a service facility of the type in Figure 6.5, he has a fixed probability b_0 of immediately leaving the facility, experiencing a zero length service time. On the other hand, there is a fixed probability a_0 that he will enter the service facility. If he enters, then he immediately proceeds to stage one. The service rate at this stage is μ_1, and the mean service time $1/\mu_1$. Upon completing service at stage one the customer proceeds to stage two according to probability a_1, or exits the facility according to probability b_1. If a customer reaches the last stage, then after receiving service he exits the facility. The mean service time of a customer is the weighted sum of the mean time spent in each stage. More precisely,

$$E(\tau) = \frac{1}{\mu_1}a_0(1-a_1) + \left[\frac{1}{\mu_1}+\frac{1}{\mu_2}\right]a_0a_1(1-a_2) + \left[\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}\right]a_0a_1a_2(1-a_3)$$

$$+ \cdots + \left[\frac{1}{\mu_1}+\frac{1}{\mu_2}+ \cdots + \frac{1}{\mu_n}\right]a_0a_1a_2\cdots a_{n-1} \quad . \qquad (6.3)$$

A little arithmetic shows that

$$E(\tau) = \frac{a_0}{\mu_1} + \frac{a_0a_1}{\mu_2} + \frac{a_0a_1a_2}{\mu_3} + \cdots + \frac{a_0a_1a_2\cdots a_{n-1}}{\mu_n} \quad . \qquad (6.4)$$

Equation (6.4) is very useful in deriving marginal steady-state probabilities. It does not appear elsewhere in the literature nor does the equation that precedes it.

In order to illustrate the procedure assume that the Laplace transform of the density function is

$$b^*(s) = \frac{s^3 + 8s^2 + 22s + 16}{4(s^3 + 5s^2 + 8s + 4)} \; .$$

Expanding this function according to Equation 6.1 results in

$$b^*(s) = \frac{1}{4} + \frac{3s^2 + 14s + 12}{4\,(s+2)^2\,(s+1)}$$

$$= \frac{1}{4} + \frac{A}{s+2} + \frac{B}{(s+2)^2} + \frac{C}{(s+2)^2(s+1)}$$

$$= \frac{1}{4} + \frac{3}{4(s+2)} + \frac{5}{4(s+2)^2} + \frac{1}{4(s+2)^2(s+1)}$$

$$= \frac{1}{4} + \frac{3}{8}\,\frac{2}{s+2} + \frac{5}{16}\,\frac{4}{(s+2)^2} + \frac{1}{16}\,\frac{4}{(s+2)^2(s+1)} \; .$$

By equating like coefficients the a's and b's can be determined, i.e.

$$b_0 = 1/4 \qquad a_0b_1 = 3/8 \qquad a_0a_1b_2 = 5/16 \qquad a_0a_1a_2 = 1/16\ ,$$

and the results are

$$a_0=3/4, \quad a_1=1/2, \quad a_2=1/6,$$
$$b_0=1/4, \quad b_1=1/2, \quad b_2=5/6.$$

The corresponding representation of the service time is depicted in Figure 6.6.

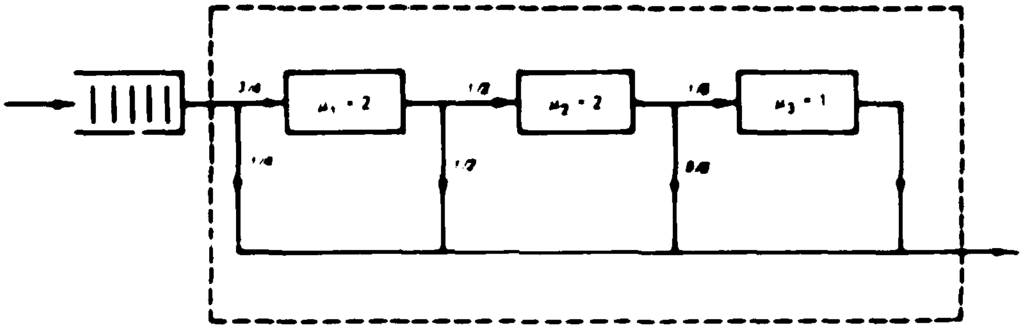


Figure 6.6 Example of Cox's Method of Exponential Stages.

A good check is to compare the mean service time calculated from the original density function with that given by Equation (6.4). This can be accomplished with the equation:

$$E[\tau] = -\left.\frac{db^*(s)}{ds}\right|_{s=0} . \qquad (6.5)$$

[KLEI75]. This equation holds for any density function and follows from the fact that the continuous time, moment generating function and the Laplace transform of the density function differ only in that the term e^{st} appears in the former and e^{-st} in the latter.

Continuing,

$$E[\tau] = \left[\frac{4(s^3+5s^2+8s+4)(3s^2+16s+22) - 4(s^3+8s^2+22s+16)(3s^2+10s+8)}{[4(s^3+5s^2+8s+4)]^2}\right]_{s=0}$$

$$= 5/8$$

and from Equation (6.4)

$$E[\tau] = \frac{3/4}{2} + \frac{(3/4)(1/2)}{2} + \frac{(3/4)(1/2)(1/6)}{1} = 5/8 \ .$$

In this example all of the poles of the Laplace transform are located on the negative real axis. However if the structure is going to be applied to any density function of a nonnegative random variable, then the poles may occur anywhere in left-half plane. This implies that the time spent in a stage may be a complex number, and that the probability of making a transition out of this stage in an infinitesimal amount of time is also complex. All of this leads to the conclusion that the simple linear balance equations of the preceding chapter now become complex equations with complex probabilities. The interpretation is that the stages are purely artificial. They are only introduced as a mathematical tool for the representation of nonexponential service times. These facts may be disturbing but they should not be. They are exactly the same principles that are used in circuit analysis. There are no complex voltages and currents in an electrical network. Complex numbers are introduced solely for the purpose of analysis. At the end of the analysis all of the results are real. The same is true here. Although the probability distribution at a fictitious stage may be complex, the distribution of customers at real service centers are real.

Before_departing the subject it should be pointed out that the problems become even worse if ordinary partial fractions are used. Not only can the time spent in a stage be complex, but the routing probabilities may be negative. Again, everything at the end turns out positive and real. For a more formal justification of complex probabilities, the reader is referred to the article by Cox referenced earlier. For a justification of negative probabilities, the reader is referred to the article by Bartlett [BART45]. Unfortunately no matter which method is used, if the coefficient of variation CV (standard deviation divided by the mean) is small, the number of stages is approximately $1/CV^2$ [COX55].

Finally, in order to incorporate Cox's method of stages into queueing networks with multiple classes, additional subscripts are required. Figure 6.7 demonstrates this. This first subscript is the service center, the second the class, and the third the stage.

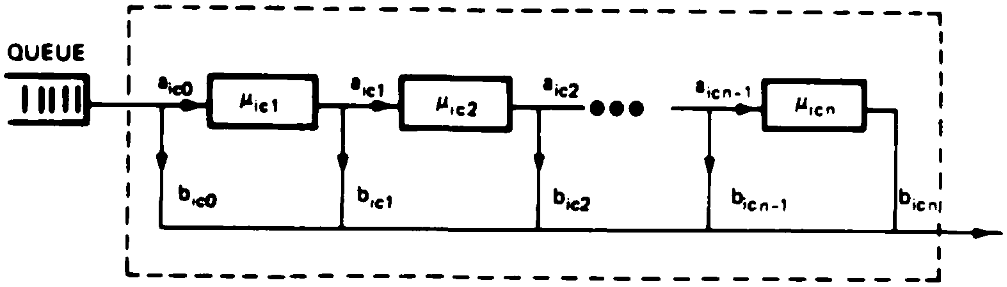


Figure 6.7 Notation Required to Incorporate Cox's Method of Exponential Stages into Queueing Networks.

6.3 Service Disciplines

In the queueing networks of the last chapter, the distribution of customers at a service center was invariant to the service discipline as long as it was work conservative. The reasons for this were that all customers were identical and all service time distributions exponential. If, however, there are customer classes associated with a service center then all customers are not identical, and the order in which customer are served plays an important role. The same is true if the service distribution is not exponential. This is because the exponential distribution is the only one that has the memoryless property. Thus, if a customers service is interrupted the probability of a transition to another state is not the same as before service was interrupted. The role that the service discipline plays should become clearer in the next section when the state space is discussed.

6.3.1 Preemptive and Nonpreemptive

If once a customer begins service he cannot be interrupted, the service discipline is said to be nonpreemptive. If on the other hand a customer can be interrupted, the discipline is said to be preemptive. If a customer that was interrupted resumes service later at the same point that he was interrupted, then the discipline is said to be preemptive resume. If the service discipline does not depend on any aspect of the customers' service demand, then the discipline is said to be service demand independent. The following is a short description of some of the more popular service disciplines:

6.3.2 First-Come-First-Service

The service discipline in which customers are served in the order of their arrival at the service center is called First Come First Serve (FCFS). It is nonpreemptive and service demand independent.

6.3.3 Priority

A priority service discipline is one in which customers are classified into types and assigned a priority according to their type. The next customer to be served is the one that has the highest priority. If more than one customer has the same priority the one that arrived first will be served first. If an arriving customer can interrupt the service of a customer with a lower priority then the discipline is called preemptive priority. If service cannot be interrupted, it is called nonpreemptive priority.

6.3.4 Round-Robin

Almost all interactive computer systems use the Round-Robin (RR) service discipline or some derivative of it. It is also referred to as time-slicing. It is defined with respect to a fixed interval of time called a quantum (or time-slice). Customers are served by a single server in first-come-first-serve order as long as their service times do not exceed the quantum. When a customer's current service time reaches the quantum, he is preempted . A preempted customer reenters the queue at the end (as if he had just arrived), and waits to receive an additional quantum of service. Each customer repeats this process until his service demand is satisfied. The advantage of this discipline is that no customer has to wait a long period of time before receiving some service.-Thus, customers with short service demands may arrive after customer with long service demands and finish ahead of them.

6.3.5 Processor-Sharing

The service discipline in which all customers receive equal and simultaneous service from a single server is called Processor Sharing (PS). That is, if there are k customers at the service center then each customer simultaneously receives service but at a rate of (1/k)th the service rate . When a new customer arrives at a service center, he immediately begins to receive service at the expense of reducing the service rate to the other customers. When a customer completes his service, the share of the server he was receiving is divided equally among all of the remaining customers. The PS service discipline cannot be actually implemented, but is an excellent approximation of the RR discipline when the quantum size is small compared to the mean service time. Analytical results are much simpler than those for RR.

6.3.6 Last-Come-First-Serve-Preemptive-Resume

A service discipline that is strictly preemptive is last-come-first-service-preemptive-resume (LCFSPR). When a new customer arrives at the service center, he interrupts the customer that is receiving service and immediately starts to receive service. When a customer finishes being served, the customer that was last interrupted resumes his service. Although this service discipline is rarely used in practice, it is included because most of the results that hold for PS also hold for LCFSPR.

6.3.7 Infinite Servers

If the number of servers at a service center is infinite or at least equal to the number of customers that can demand service simultaneously, then the service center is said to have an infinite server (IS) service discipline. Customers always begin receiving service immediately upon arrival, therefore there is no service order or waiting line. Also, there is never contention for a server. The servers usually do not represent physical resources. Service centers of this type are used almost exclusively to represent delays that occur in real processes. It is always possible to coalesce IS service centers into a single service center by incorporating new classes.

6.4 The State Space

As previously stated, in order that a process be a Markov process, it is necessary that the state of the process summarize all pertinent past history. For a service center with multiple classes and/or nonexponential service times, the history that must be contained in the state depends on the service discipline. For example, if there are multiple classes at a service center and the service discipline is FCFS, then the state must contain the order and class of customer in the queue and service facility. If the service time is nonexponential and the service discipline is processor sharing, then the state must contain the stage of service that each customer is in. If there are multiple classes and nonexponential service times and the service discipline is preemptive, then the state must contain the order in which customers are to be served, the class of each customer, and the stage of service that a customer was in before he was preempted.

In order to keep the notation manageable, the state space will be defined only for the class of queueing networks that have product form solutions. It makes little sense to do otherwise, since at the present these are the only networks for which exact solutions can be obtained. Exceptions are closed networks with a small number of service centers and customers (See Chapter 7, Section 7.4). The notation and many of the results in this chapters are from the article 'Open, Closed, and Mixed Networks of Queues with Different Classes of Customers' by Baskett, Chandy, Muntz, and Palacious [BASK75].

Service centers will be referred to as types FCFS/1/, PS, LCFSPR, or IS according to the following:

FCFS/1/ - The service center has a single server and the service discipline is first-come-first-serve (FCFS). In addition all customers must have the same service time distribution and the distribution must be exponential. The service rate may not depend on the number of customers at the service center (later in this chapter this restriction will be removed and multiple servers allowed).

PS - The service center has a single server and the service discipline is processor sharing (i.e., when there are k customers each is receiving service simultaneously at a rate of 1/k seconds of service per second). Each class of customers may have a distinct service time distribution, however all density functions must have rational Laplace transforms.

LCFSPR - The service center has a single server and the service discipline is last-come-first-serve-preemptive-resume (LCFSPR). Each class of customers receiving service at this center may have a distinct service time distribution, however all density functions must have rational Laplace transforms.

IS - The number of servers at this type of service center is infinite (or at least equal to the number of customers which can be demanding service simultaneously at this center). Each class of customer receiving service at this center may have a distinct service time distribution, however all density functions must have rational Laplace transforms. Service centers of this type are said to have an infinite server (IS) service discipline.

Any queueing network composed of service centers of these types has a product form solution. The next section is concerned with the justification of this. The necessary conditions in order for a network to have a product form solution are : (1) the service discipline is FCFS and all customers have the same service time distribution regardless of class, or (2) the service discipline at the service center must be such that every customer starts to receive some service immediately upon arriving [CHAN77] [CHAN83]. Service center types PS, LCFSPR, and IS satisfy the second condition. Although there may be other service disciplines which satisfy the second condition they are of little practical significance and will not be discussed here.

The state of the network with N service centers and C classes is a vector $(x_1,x_2,\ldots,x_N)$ where x_i represents the conditions prevailing at service center i. The representation of x_i depends on the type of service center i.

If service center i is type FCFS/1/, then

$$x_i = (x_{i1},x_{i2},\ldots,x_{ik_i})$$

where

k_i = the number of customers in service center i

and

x_{ij} = the class of the customer jth in FCFS order. (6.6)

If service center i is type PS or IS, then

$$x_i = (u_{i1},u_{i2},\ldots,u_{iC})$$

where

$$u_{ic} = (\alpha_{1c},\alpha_{2c},\ldots,\alpha_{z_{ic}c})$$

and

α_{nc} = the number of class c customer in the nth stage of service

and

z_{ic} = total number of stages for a class c customer. (6.7)

If service center i is type LCFSPR then,

$$x_i = ((c_1,\alpha_1),(c_2,\alpha_2),\ldots,(c_{k_i},\alpha_{k_i}))$$

where

k_i = the number of customers in service center i

and

(c_j,α_j) = the class and stage of the jth customer in LCFSPR order. (6.8)

6.5 The Steady-State Solution

For a network with C classes and N service centers of type FCFS/1/$\overline{\ ,\ }$PS, LCFSPR, and IS the steady-state probabilities are given by:

$$P(x_1,x_2,\ldots,x_N) = \lambda(S(K))\ \frac{f_1(x_1)\ f_2(x_2)\ \ldots\ f_N(x_N)}{G}\ , \qquad (6.9)$$

where G is normalizing constant chosen to make the steady-state probabilities sum to one, $\lambda(S(K))$ is a function that depends on the arrival process, and each f_i is a function that depends on the type of service center i.

If the network is closed, then $\lambda(S(K)) = 1$. (6.10)

If the network is open and there is only one chain, then

$$\lambda(S(K)) = \prod_{a=0}^{K-1} \lambda(a)\ ,$$

where K equals the number of customers in the network, and $\lambda(a)$ is the mean arrival rate when the network has 'a' customers. (6.11)

If the network is open and there are J chains, then

$$\lambda(S(K)) = \prod_{j=1}^{J}\ \prod_{a=0}^{K_j} \lambda_j(a)\ ,$$

where K_j equals the number of customers in chain j, and $\lambda_j(a)$ is the mean arrival rate to chain j when it has 'a' customers. (6.12)

In order to simplify the equations for $f_i(x_i)$ let

$$A_{icm} = \prod_{j=0}^{m-1} a_{icj} \quad . \tag{6.13}$$

If service center i is type FCFS/1/, then

$$f_i(x_i) = (1/\mu_i)^{k_i} \prod_{j=1}^{k_i} e_{ix_{ij}} \quad . \tag{6.14}$$

If service center i is type PS, then

$$f_i(x_i) = k_i! \prod_{c=1}^{C} \prod_{m=1}^{x_{ic}} ([e_{ic} A_{icm}/\mu_{icm}]^{a_{icm}} (1/a_{icm}!)) \quad . \tag{6.15}$$

If service center i is type LCFSPR, then

$$f_i(x_i) = \prod_{j=1}^{k_i} [e_{ic_j} A_{ic_jm_j} (1/\mu_{ic_jm_j})] \quad . \tag{6.16}$$

If service center i is type IS, then

$$f_i(x_i) = \prod_{c=1}^{C} \prod_{m=1}^{x_{ic}} ([e_{ic} A_{icm}/\mu_{icm}]^{a_{icm}} (1/a_{icm}!)) \quad . \tag{6.17}$$

All empty product terms are assigned the value of +1.

These statements are presented as a theorem in the paper by Baskett referenced earlier. The following is taken directly from that paper : ' The theorem is proved by checking that the independent (local) balance equations are satisfied. In every case for which these results apply the independent balance equations reduces to the defining quations for the e_{ic}. These two sentences are the only justification given. Several months were spent trying to prove by local balance that the theorem holds for the general network. Unfortunately, it could only be shown to hold for specific networks (no general types other than those in the last chapter). There are just too many degrees of freedom. The generalized network contains an arbitrary, but finite number of service centers each of which may be types FCFS/1/, PS, LCFSPR, or IS. In addition, there is an arbitrary but finite number of classes associated with each service center, and for types PS, LCFSPR, and IS the service time for each class may be represented by an arbitrary but finite number of exponential stages. The network may be either open, closed, or mixed, and may contain an arbitrary but finite number of chains. If the network is open or mixed, then each open chain may have its own arrival process which may depend on the number of customers in the chain.

The conclusion reached is that the theorem holds for the generalized network, but it cannot be proven to do so, at least not by the technique of local balance. Some other disappointing and disturbing facts are : (1) the equations for $f_i(x_i)$ types PS, LCFSPR, and IS are all incorrect as stated in the original paper (indices and subscript errors), (2) the equations do not appear elsewhere in open literature, and (3) there are no examples in open literature showing the theorem holds for a specific network. These statements are based on over three years of research in this area. The reason for bringing out these deficiencies is that all of them will at least be partially addressed here.

6.5.1 A New Justification for Networks with Service Centers of Type FCFS, PS, and IS

Figure 6.8 depicts an arbitrary network, and Figure 6.9 a blow-up view of service center 3. Observe that there are two customer classes at service center 3 and that the service time distribution of both classes are represented by Cox's method of exponential stages.

For the moment assume that service center 3 is type IS. That is there are an infinite number of servers at service center 3, and there is never a waiting line or queue. In addition all customers are receiving service simultaneously, and no customer or class of customers affects the service of any other customer. Thus, each stage behaves as an independent service center of type IS with an exponential distributed service time. It follows that the network in Figure 6.10 is equivalent to the network in Figure 6.8. Furthermore, the service centers and routing probabilities can be relabled as in Figure 6.11 to eliminate all class and stage subscripts. Although it was not explicitly stated in the last chapter, a FCFS service center with an exponential service time distribution and a service rate of

$$\mu_i(k_i) = k_i \ \mu_i \quad \text{for } k_i > 0 \ ,$$

is identical to an IS service service with the same service time distribution. Thus, for the network in Figure 6.11

$$\prod_{i=3}^{7} f_i(k_i) = \left[\frac{e_3}{\mu_3}\right]^{k_3} \frac{1}{k_3!} \quad \left[\frac{e_4}{\mu_4}\right]^{k_4} \frac{1}{k_4!} \quad \left[\frac{e_5}{\mu_5}\right]^{k_5} \frac{1}{k_5!} \quad \left[\frac{e_6}{\mu_6}\right]^{k_6} \frac{1}{k_6!} \quad \left[\frac{e_7}{\mu_7}\right]^{k_7} \frac{1}{k_7!}$$

where

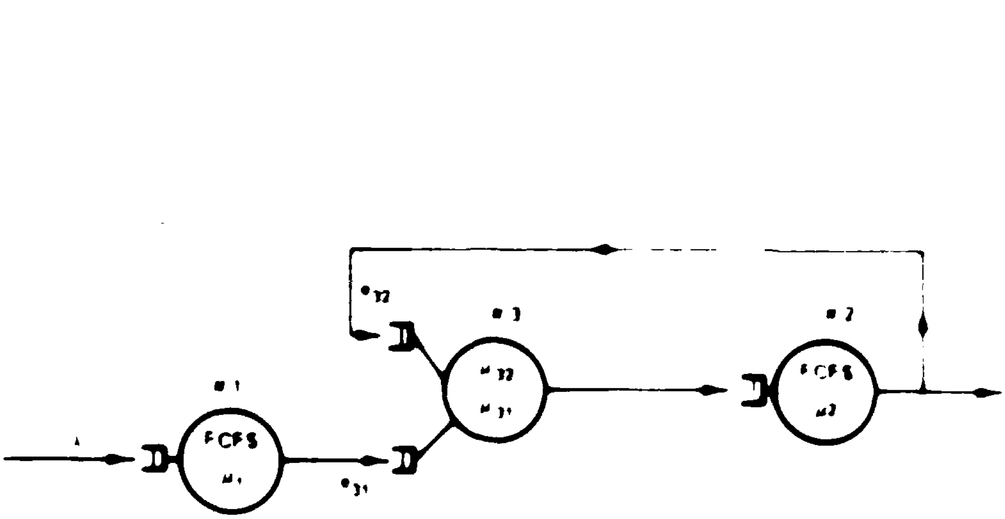


Figure 6.8 An Arbitrary Queueing Network.

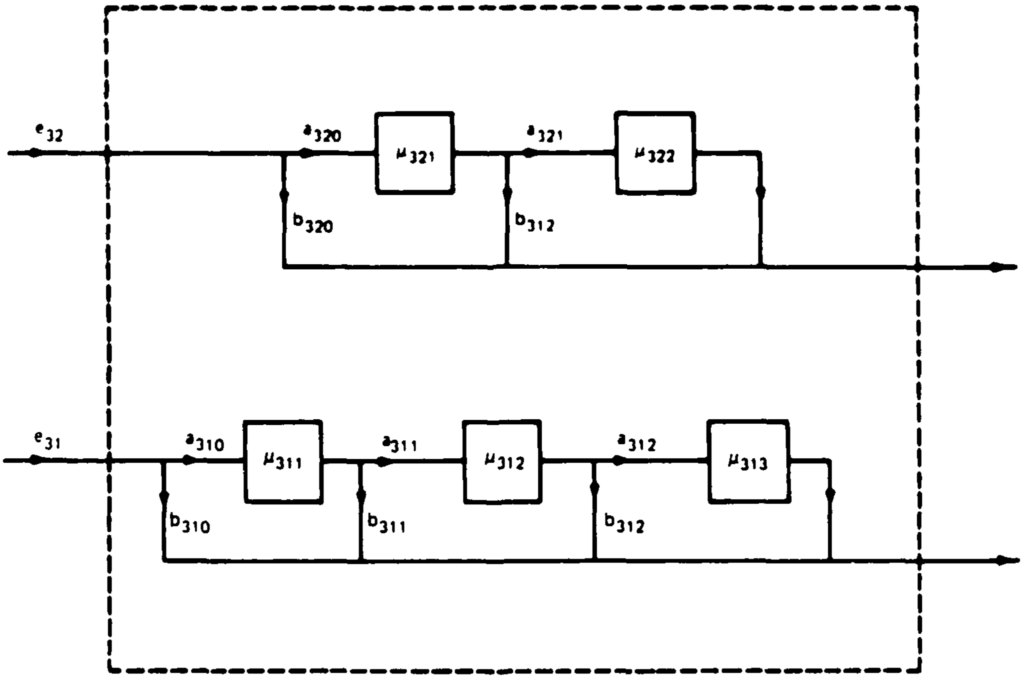


Figure 6.9 Blow-Up View of Service Center #3.

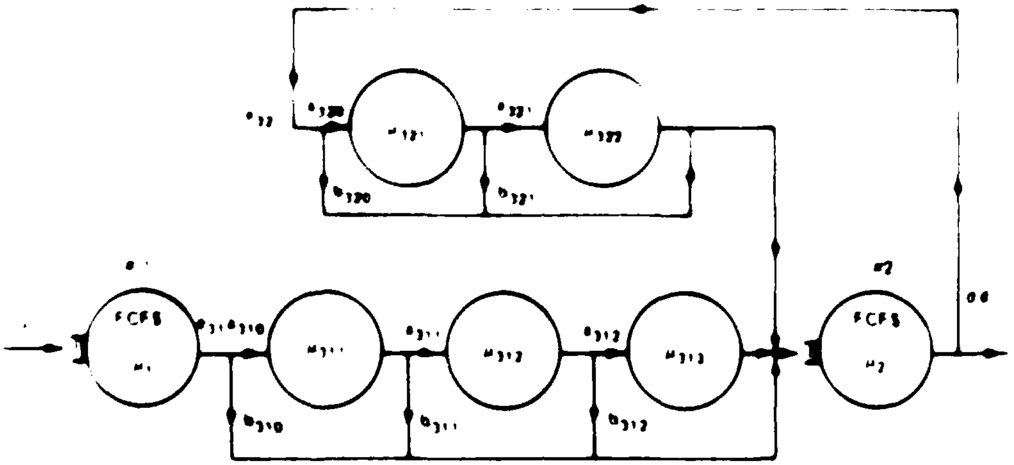


Figure 6.10 Equivalent Network of Figure 6.8.

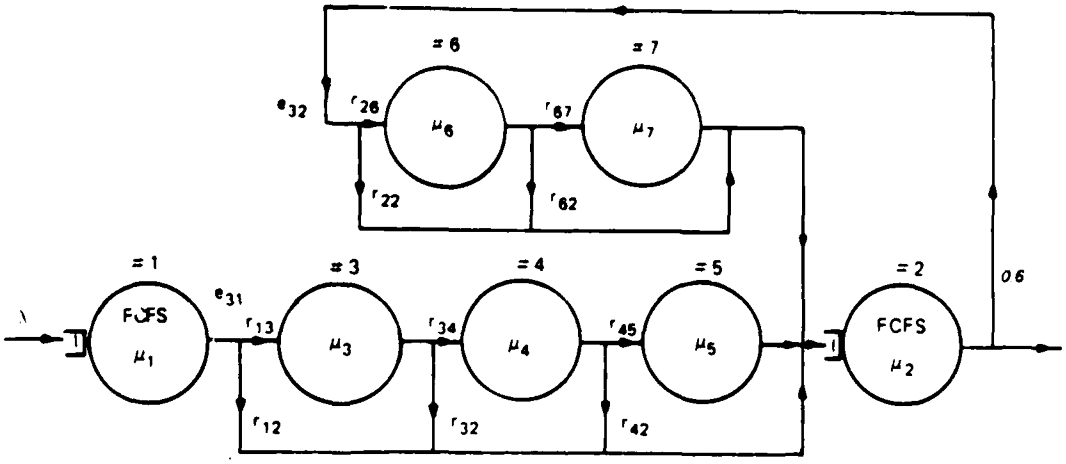


Figure 6.11 Equivalent Network of Figure 6.9.

$$f_i(k_i) = \prod_{a=1}^{k_i} e_i/\mu_i(a) \ .$$

(Note that the equation for $f_i(k_i)$ is from chapter 5, and that load dependent service rates will not be covered until later in this chapter). Solving for the relative throughputs in terms of the class throughputs and routing probabilities yields:

$$\prod_{i=3}^{7} f_i(k_i) = \left[\frac{e_{31}\ r_{13}}{\mu_3}\right]^{k_3} \frac{1}{k_3!} \quad \left[\frac{e_{31}\ r_{13}\ 34}{\mu_4}\right]^{k_4} \frac{1}{k_4!}$$

$$\left[\frac{e_{31}\ r_{13}\ r_{34}\ r_{45}}{\mu_5}\right]^{k_5} \frac{1}{k_5\ !} \quad \left[\frac{e_{32}\ r_{26}}{\mu_6}\right]^{6^k} \frac{1}{k_6!}$$

$$\left[\frac{e_{32}\ r_{26}\ r_{67}}{\mu_7}\right]^{k_7} \frac{1}{k_7!} \ .$$

Similarly, $f_3(x_3)$ in Figure 6.9 and the product of the $f_i(x_i)$ over the equivalent set of service centers in Figure 6.10 is

$$f_3(x_3) = \left[\frac{e_{31}\ a_{310}}{\mu_{311}}\right]^{\alpha_{311}} \frac{1}{\alpha_{311}!} \quad \left[\frac{e_{31}\ a_{310}\ a_{311}}{\mu_{312}}\right]^{\alpha_{312}} \frac{1}{\alpha_{312}!}$$

$$\left[\frac{e_{31}\ a_{310}\ a_{311}\ a_{312}}{\mu_{313}}\right]^{\alpha_{313}} \frac{1}{\alpha_{313}!} \quad \left[\frac{e_{32}\ a_{320}}{\mu_{321}}\right]^{\alpha_{321}} \frac{1}{\alpha_{321}!}$$

$$\left[\frac{e_{32}\ a_{320}\ a_{321}}{\mu_{322}}\right]^{\alpha_{322}} \frac{1}{\alpha_{322}!} \ .$$

Now consider the case in which the service discipline at service center 3 is processor sharing. There is also never a waiting line or queue at this type of service center. More precisely, the server is shared equally among all customers. Thus, the only difference between service center types IS and PS is that a customer in the first receives one second of service per second and a customer in latter receives $1/k_i$ seconds of service per second, where k_i is the total number of customer in the PS service center. Thus, it follows that service center 3 can be expanded into a subnetwork of service centers just as was done for the IS case with the exception that the service rate becomes

$$\mu_i(k_i)\ Z(K_I) = \frac{k_i\ \mu_i}{k_I} \qquad \text{for } k_i>0\ ,$$

where i is an arbitrary service center in the subnetwork, and K_I the total number of customers in the subnetwork. It follows from Chapter 5, Section 5.4.5 that

$$\prod_{i=3}^{7} f_i(k_i) := \prod_{a=1}^{K_I} (1/Z_I(a)) \prod_{i=3}^{7} f_i(x_i) = K_I! \prod_{i=3}^{7} f_i(x_i)\ ,$$

which accounts for the fact that the only difference in $f_i(x_i)$ for type IS and PS service center is that the latter is preceded by k_i (the symbol := is an assignment operator).

The conclusion is that any network composed of FCFS, PS, and IS service centers can be mapped into an equivalent network by letting the stages grow to full service centers. Hence, the proofs in the previous chapter are sufficient to cover a large class of networks, but not all of those considered by Baskett.

6.5.2 Local Balance

Local balance was described in the last chapter and is easily extended to networks with classes and nonexponential service times. It equates the rate of flow into a state due to a class c customer entering a stage of service, to the flow out of that state due to a class c customer leaving that stage of service. From the description of local balance, it is easily seen that each global balance equation (equates total rate of flow into a state to total rate of flow out of a state) is a sum of local balance equations. Therefore, the solution to the local balance equations satisfies the global balance equations.

Some insight can be gained by recalling the origins of global and local balance equations. Global balance equations are derived from the fact that a queueing network is a multidimensional, birth and death, Markov process. It is the solution of the global balance equations that is important. However if the network is open, there is no way to solve these equations mathematically (there are always more unknowns than equations). The only way to obtain a solution is to guess. One way of guessing is to assume local balance. That is, local balance is an assumption which may or may not be true. If the assumption is false then the local balance equations will be inconsistent.

The technique and power of local balance will be demonstrated by an example. Consider the problem of finding the steady-state probabilities for the network in Figure 6.12. The service discipline is processor sharing and there are two customer classes. Class 1 and class 2 customers arrive from Poisson sources with mean rates of λ_1 and λ_2

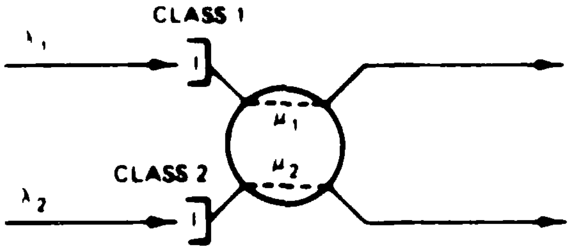


Figure 6.12 Type PS Service Center with Two Classes.

respectively (the two sources could be combined, but this will not be done here). The service times of class 1 and class 2 customers are exponentials distributed with means $1/\mu_1$ and $1/\mu_2$ respectively. Since the service discipline is processor sharing, the service'.rates depend on the number of customers in the service center. The service rate of a class 1 customer is $\mu_1/(k_1+k_2)$, where k_1 equals the number of class 1 customers in the service center and k_2 the number of class 2 customers. Similarly, the service rate of a class 2 customer is $\mu_2/(k_1+k_2)$. The probability that a class 1 customer departs the service center in infinitesimal time h is $[h\ k_1\ \mu_1\ /\ (k_1+k_2)]$. Similarly, for a class 2 customer the probability is $[h\ k_2\ \mu_2\ /\ (k_1+k_2)]$. Let the steady-state probabilities be represented by $P(k_1,k_2)$, where k_1 and k_2 are the number of class 1 and class 2 customers. The state-transition-rate diagram of the network is depicted in Figure 6.13.

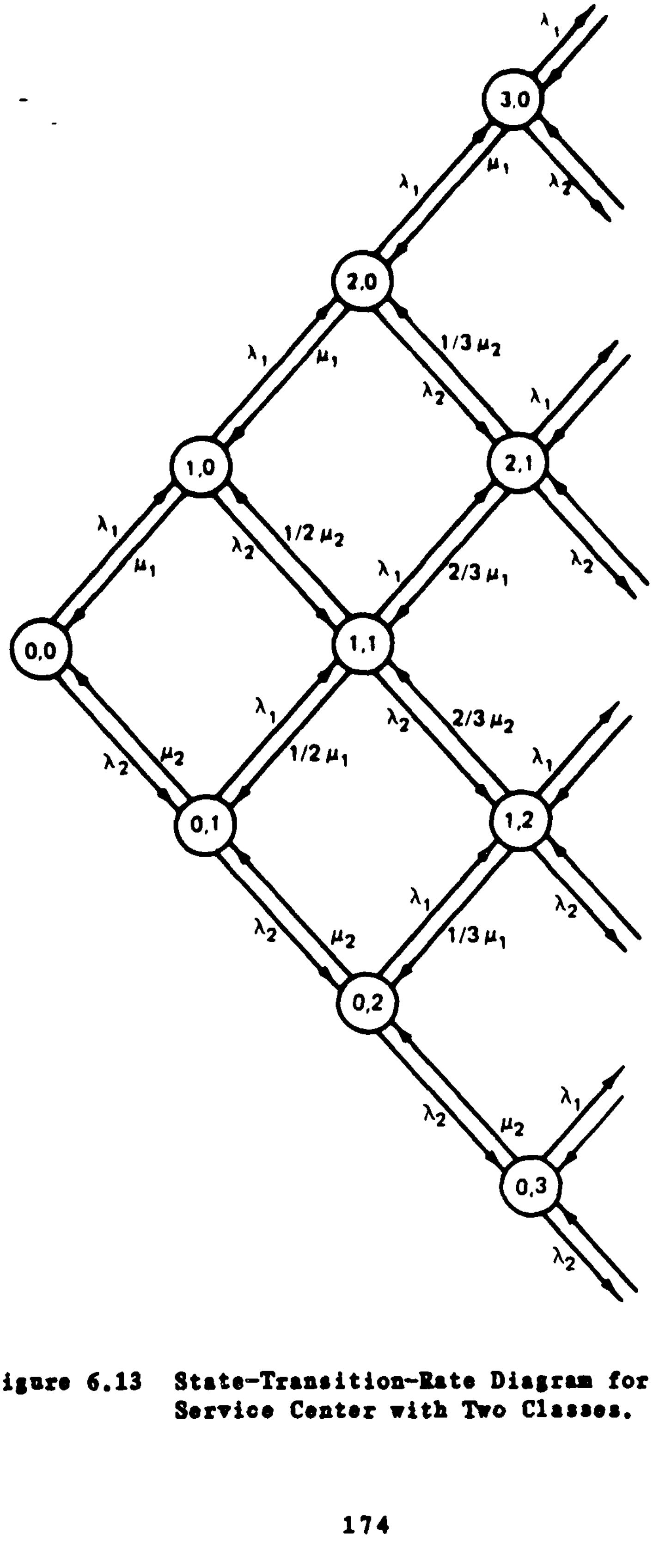


Figure 6.13 State-Transition-Rate Diagram for Type PS Service Center with Two Classes.

The global balance equations for states with two or fewer customers are:

$$(\lambda_1+\lambda_2)P(0,0) = \mu_1 P(1,0) + \mu_2 P(0,1)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,0) = \mu_1 P(2,0) + \frac{1}{2}\mu_2 P(1,1) + \lambda_1 P(0)$$

$$(\lambda_1+\lambda_2+\mu_2)P(0,1) = \frac{1}{2}\mu_1 P(1,1) + \mu_2 P(0,2) + \lambda_2 P(0)$$

$$(\lambda_1+\lambda_2+\mu_1)P(2,0) = \mu_1 P(3,0) + \frac{1}{3}\mu_2 P(2,1) + \lambda_1 P(1,0)$$

$$(\lambda_1+\lambda_2+ \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2)P(1,1) = \frac{2}{3}\mu_1 P(2,1) + \frac{2}{3}\mu_2 P(1,2) + \lambda_1 P(0,1) + \lambda_2 P(1,0)$$

$$(\lambda_1+\lambda_2+\mu_2)P(0,2) = \frac{1}{3}\mu_1 P(1,2) + \mu_2 P(0,3) + \lambda_2 P(0,1) .$$

Notice that there are six equations and ten unknowns. No matter how many equations are written out there will always be more equations than unknowns! The corresponding local balance equations are:

$$\lambda_1 P(0,0) = \mu_1 P(1,0)$$
$$\lambda_2 P(0,0) = \mu_2 P(0,1)$$

$$\lambda_1 P(1,0) = \mu_1 P(2,0)$$
$$\lambda_2 P(1,0) = \frac{1}{2}\mu_2 P(1,1)$$
$$\mu_1 P(1,0) = \lambda_1 P(0)$$

$$\lambda_1 P(0,1) = \frac{1}{2}\mu_1 P(1,1)$$
$$\lambda_2 P(0,1) = \mu_2 P(0,2)$$
$$\mu_1 P(0,1) = \lambda_2 P(0)$$

$$\lambda_1 P(2,0) = \mu_1 P(3,0)$$

$$\lambda_2 P(2,0) = \frac{1}{3}\mu_2 P(2,1)$$

$$\mu_1 P(2,0) = \lambda_1 P(1,0)$$

$$\lambda_1 P(1,1) = \frac{2}{3}\mu_1 P(2,1)$$

$$\lambda_2 P(1,1) = \frac{2}{3}\mu_2 P(1,2)$$

$$\frac{1}{2}\mu_1 P(1,1) = \lambda_1 P(0,1)$$

$$\frac{1}{2}\mu_2 P(1,1) = \lambda_2 P(1,0)$$

$$\lambda_1 P(0,2) = \frac{1}{3}\mu_1 P(1,2)$$

$$\lambda_2 P(0,2) = \mu_2 P(0,3)$$

$$\mu_1 P(0,2) = \lambda_2 P(0,1) \ .$$

Solving these equations in terms of P(0,0) yields:

$$P(1,0) = (\lambda_1/\mu_1)\ P(0,0)$$

$$P(0,1) = (\lambda_2/\mu_2)\ P(0,0)$$

$$P(2,0) = (\lambda_1/\mu_1)^2\ P(0,0)$$

$$P(1,1) = 2\ (\lambda_1/\mu_1)\ (\lambda_2/\mu_2)\ P(0,0)$$

$$P(0,2) = (\lambda_2/\mu_2)^2\ P(0,0)$$

$$P(3,0) = (\lambda_1/\mu_1)^3\ P(0,0)$$

$$P(2,1) = 3\ (\lambda_1/\mu_1)^2\ (\lambda_2/\mu_2)\ P(0,0)$$

$$P(1,2) = 3\ (\lambda_1/\mu_1)\ (\lambda_2/\mu_2)^2\ P(0,0)$$

$$P(0,3) = (\lambda_2/\mu_2)^3\ P(0,0) \ .$$

Although it may be a little difficult to see without solving more equations the form of the solution is :

$$P(k_1,k_2) = [\lambda_1^{k_1}\ \lambda_2^{k_2}]\ [(k_1+k_2)!\ (1/\mu_1)^{k_1}(1/k_1!)\ (1/\mu_2)^{k_2}(1/k_2!)]\ P(0,0).$$

Notice that the first term in brackets corresponds to the equation for $\lambda(S(K))$ with two chains. The second term in brackets corresponds to the equation for $f_i(x_i)$ (type PS), where $e_{i1} = e_{i2} = 1$. Obviously G equal $1/P(0,0)$.

The power of local balance is : answers can be obtained, and it is only necessary to guess at the general form of the solution. In order to show the equations for $f_i(x_i)$ are valid for other types of service centers, this same problem is worked in Appendix A for FCFS and LCFSPR service disciplines. In addition, Appendix A also contains an example of a service center with two exponential stages and LCFSPR service discipline. As previously stated these are the only known examples demonstrating that the steady-state equations in this section are valid.

Unfortunately, only a small subset of service centers have local balance. For example, if the service discipline of the network in Figure 6.12 is changed to nonpreemptive priority, local balance is not applicable. This can easily be seen by writing the global balance equations. Let $(x_1,x_2,\ldots,x_k)$ represent the state of the network, where x_1 is the class of the customer currently being served, x_2 the class of the customer to be served next, x_3 the class of the customer after x_2, etc. The global balance equations for all states with two or fewer customers are:

$$(\lambda_1+\lambda_2)P(0) = \mu_1P(1) + \mu_2P(2)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1) = \mu_1P(1,1) + \mu_2P(2,1) + \lambda_1P(0)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2) = \mu_1P(1,2) + \mu_2P(2,2) + \lambda_2P(0)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,1) = \mu_1P(1,1,1) + \mu_2P(2,1,1) + \lambda_1P(1)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,2) = \mu_1P(1,1,2) + \mu_2P(2,1,2) + \lambda_2P(1)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,1) = \lambda_1P(2)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,2) = \mu_1P(1,2,2) + \mu_2P(2,2,2) + \lambda_2P(2) .$$

To show that local balance is not applicable all one needs to do is examine the sixth equation. There are three ways to depart state (2,1) and only one way to enter it. Thus, local balance cannot apply, nor can it be extended to do so. This counter example proves conclusively that the same techniques that are used to solve networks with type FCFS/1/, PS, LCFSPR, and IS service centers do not apply if the service discipline is nonpreemptive priority. In every single case there will always be some network state (usually many) in which there are more ways to depart a state than enter it. This is not surprising since it does not meet the necessary conditions for a product form solution.

It is also proven in Appendix A that local balance is not applicable if the service discipline is FCFS and customers do not have the same service rate, μ. In this case local balance equations can be written, but they are inconsistent.

6.5.3 Marginal Distributions

The more detailed states in section 6.5 are necessary to derive the steady-state probabilities. In this section marginal distributions are obtained by aggregating states. Let the aggregate system state be the number of customers of each class in each service center. More formally the aggregate system state is defined to be the vector $(y_1,y_2,\ldots,y_N)$, where $y_i = (k_{i1},k_{i2},\ldots,k_{iC})$ and k_{ic} is the number of customers of class c in service center i. Also, let $1/\mu_{ic}$ be the mean service time of class c customer at service center i. The steady-state aggregate probabilities are given by

$$P(y_1,y_2,\ldots,y_N) = \lambda(S(K)) \, \frac{g_1(y_1)\; g_2(y_2)\; \cdots \; g_N(y_N)}{G} \quad .$$

where

$$g_i(y_i) = \begin{cases} k_i! \displaystyle\prod_{c=1}^{C} (1/k_{ic}!)\; (e_{ic}/\mu_{ic})^{k_{ic}} \; , & \text{for } i \text{ FCFS/1/, PS, or LCFSPR,} \\ \displaystyle\prod_{c=1}^{C} (1/k_{ic}!)\; (e_{ic}/\mu_{ic})^{k_{ic}}, & \text{for } i \text{ IS.} \end{cases} \tag{6.18}$$

The expressions for $g_i(y_i)$ are derived by summing $f_i(x_i)$ over all x_i with $k_{i1},k_{i2},\ldots,k_{iC}$ fixed. The multinomial theorem and Equation 6.2 are useful (an example will be given later). Note that for type FCFS/1/ service centers the μ_i has been moved inside of the summation and changed to μ_{ic}. This was done both to simplify the notation and to emphasize the similarities between FCFS/1/, PS, and LCFSPR. It is however required that for FCFS/1/

$$\mu_{i1} = \mu_{i2} = \cdots = \mu_{iC} \; . \tag{6.19}$$

The fact that the number of customers in a closed chain is constant suggests eliminating class distinction and distinguishing customers according to chains. More precisely let the network state be $(z_1, z_2, \ldots, z_N)$ where $z_i(k_{i1}, k_{i2}, \ldots, k_{iJ})$ and k_{ij} is the number of chain j customers at service center i. It follows from the multinomial theorem that

$$P(z_1, z_2, \ldots, z_N) = \lambda(S(K)) \frac{w_1(z_1)\ w_2(z_2)\ \ldots\ w_N(z_N)}{G}$$

where

$$w_i(z_i) = \begin{cases} k_i! \prod_{j=1}^{J} (1/k_{ij}!)\ (e_{ij}/\mu_{ij})^{k_{ij}}\ , & \text{for } i \text{ FCFS/1, PS, or LCFSPR,} \\ \prod_{j=1}^{J} (1/k_{ij}!)\ (e_{ij}/\mu_{ij})^{k_{ij}}, & \text{for } i \text{ IS,} \end{cases}$$

$$k_{ij} = \sum_{c \text{ in } j} k_{ic} = \text{number of chain } j \text{ customers at service center } i,$$

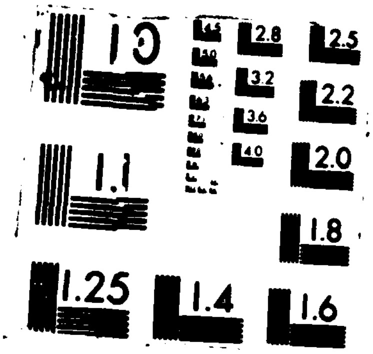
$$e_{ij} = \sum_{c \text{ in } j} e_{ic} = \text{the relative throughput of a chain } j \text{ customer through service center } i,$$

$$1/\mu_{ij} = (1/e_{ij}) \sum_{c \text{ in } j} (e_{ic}/\mu_{ic}) = \text{mean service time of a chain } j \text{ customer at service center } i.$$

Observe that the equation for $w_i(z_i)$ is isomorphic to the equation for $g_i(y_i)$. That is, class parameters are simply replaced by chain parameters. Also note that if there is only one class per chain

AD-A182 994 NETWORKS OF MARKOVIAN QUEUES(U) ARMY ENGINEER WATERWAYS 3/4
EXPERIMENT STATION VICKSBURG MS INSTRUMENTATION
SERVICES DIV R A FRANCO MAY 87 WES/TR/O-87-2

UNCLASSIFIED F/G 12/3 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

$w_i(z_i) = g_i(y_i)$. The advantage of this aggregate state over the previous one is that the number of feasible network states has been significantly reduced. It follows that since the purpose of G is to force the sum of the probabilities over all feasible networks states to one that it is much easier to determine G using this aggregate state. If one requires probability distributions by class the value of G can then be substituted into the previous set of equations.

A further simplication is possible by defining the aggregate state as the total number of customers k_i in the ith service center (eliminating class and chain distinctions). That is, let the state be $(k_1, k_2, \ldots, k_N)$ where k_i equal the number of customers at service center i. If follows that the steady-state probabilities are given by :

$$P(k_1, k_2, \ldots, k_N) = \lambda(S(K)) \frac{h_1(k_1)\ h_2(k_2)\ \ldots\ h_N(k_N)}{G}$$

where

$$h_i(k_i) = \begin{cases} \left[\sum\limits_c (e_{ic}/\mu_{ic}) \right]^{k_i} & \text{if } i \text{ FCFS/1/, PS, or LCFSPR} \\ (1/k_i!) \left[\sum\limits_c (e_{ic}/\mu_{ic}) \right]^{k_i} & \text{if } i \text{ IS ,} \end{cases} \qquad (6.21)$$

where the sum is over all customer classes which may enter service center i.

Although this simplification is valid for both open and closed networks, algorithms to calculate the normalizing constant do not use it (closed networks). They rely on the fact that number of customers in each chain is constant. On the other hand, if the network is open and the mean arrival rate is constant, then a closed form expression for the normalizing constant can be obtained with the aid of these equations. This will be done in the next section. In contrast to the earlier equations this set of marginal steady-state equations appear in several places in the literature [KLEI76] [BRUE80] [SAUE81] [LAVE83].

As promised earlier the material in the section will be illustrated by an example. Assume that the service center in Figure 6.14 is type PS.

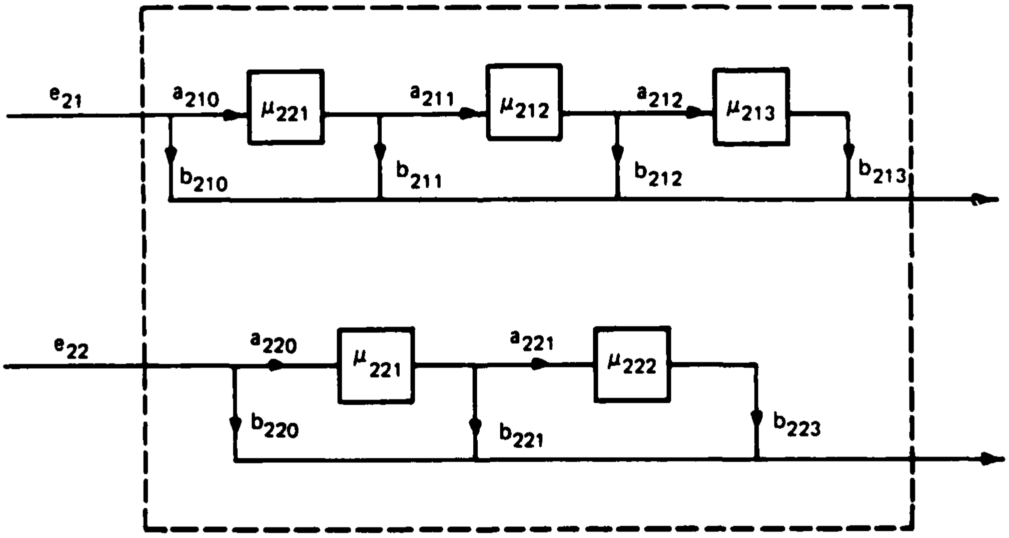


Figure 6.14 PS Type Service Center with Two Customer Classes.

For this service center,

$$f_2(\bar{x}_2) = k_2! \left[\frac{e_{21}\ a_{210}}{\mu_{211}}\right]^{\alpha_{211}} \frac{1}{\alpha_{211}!} \left[\frac{e_{21}\ a_{210}\ a_{211}}{\mu_{211}}\right]^{\alpha_{212}} \frac{1}{\alpha_{212}!}$$

$$\left[\frac{e_{21}\ a_{210}\ a_{211}\ a_{212}}{\mu_{213}}\right]^{\alpha_{213}} \frac{1}{\alpha_{213}!} \left[\frac{e_{22}\ a_{220}}{\mu_{221}}\right]^{\alpha_{221}} \frac{1}{\alpha_{221}!}$$

$$\left[\frac{e_{22}\ a_{220}\ a_{221}}{\mu_{222}}\right]^{\alpha_{222}} \frac{1}{\alpha_{222}!} \ .$$

Summing over all α_{211}, α_{212}, α_{213}, α_{221}, and α_{222} such that $\alpha_{211}+\alpha_{212}+\alpha_{213}=k_{21}$ and $\alpha_{221}+\alpha_{222}=k_{22}$ results in:

$$g_2(y_2) = k_2! \ \{ \sum_{\alpha_{211}+\alpha_{212}+\alpha_{213}=k_{21}} \left[\frac{e_{21}\ a_{210}}{\mu_{211}}\right]^{\alpha_{211}} \frac{1}{\alpha_{211}!} \ .$$

$$\left[\frac{e_{21}\ a_{210}\ a_{211}}{\mu_{211}}\right]^{\alpha_{212}} \frac{1}{\alpha_{212}!} \left[\frac{e_{21}\ a_{210}\ a_{211}\ a_{212}}{\mu_{213}}\right]^{\alpha_{213}} \frac{1}{\alpha_{213}!}$$

$$\sum_{\alpha_{221}+\alpha_{222}=k_{22}} \left[\frac{e_{22}\ a_{220}}{\mu_{221}}\right]^{\alpha_{221}} \frac{1}{\alpha_{221}!} \left[\frac{e_{22}\ a_{220}\ a_{221}}{\mu_{222}}\right]^{\alpha_{222}} \frac{1}{\alpha_{222}!} \ \}$$

$$= \frac{k_2!}{k_{21}!} \left[\frac{e_{21}\ a_{210}}{\mu_{211}} + \frac{e_{21}\ a_{210}\ a_{211}}{\mu_{212}} + \frac{e_{21}\ a_{210}\ a_{211}\ a_{212}}{\mu_{213}} \right]^{k_{21}}$$

$$\frac{k_2!}{k_{22}!} \left[\frac{e_{22}\ a_{220}}{\mu_{221}} + \frac{e_{22}\ a_{220}\ a_{221}}{\mu_{222}} \right]^{k_{22}} \ .$$

The last expression is obtained by multiplying and dividing the first summation by k_{21} and the second by k_{22}, and then recognizing the summations as special cases of the multinomial theorem. From equation 6.2 the mean service times of a class 1 and class 2 customer are respectively :

$$\frac{1}{\mu_{21}} = \frac{a_{210}}{\mu_{211}} + \frac{a_{210}\ a_{211}}{\mu_{212}} + \frac{a_{210}\ a_{211}\ a_{212}}{\mu_{213}}$$

$$\frac{1}{\mu_{22}} = \frac{a_{220}}{\mu_{221}} + \frac{a_{220}\ a_{221}}{\mu_{222}} \quad .$$

Finally, substituting these expression into the last equation for $g_2(y_2)$ yields

$$g_2(y_2) = k_2! \quad \left[\frac{1}{k_{21}!} \left[\frac{e_{21}}{\mu_{21}} \right]^{k_{21}} \frac{1}{k_{22}!} \left[\frac{e_{22}}{\mu_{22}} \right]^{k_{22}} \right] .$$

Now assuming that class 1 and 2 both belong to chain j, and summing over all k_{12} and k_{22} such that $k_{12}+k_{22}=k_2=k_j$ yields $w_2(z_2)$. More precisely,

$$w_2(z_2) = \sum_{k_{21}+k_{22}=k_2} k_2! \quad \left[\frac{1}{k_{21}!} \left[\frac{e_{21}}{\mu_{21}} \right]^{k_{21}} \frac{1}{k_{22}!} \left[\frac{e_{22}}{\mu_{22}} \right]^{k_{22}} \right]$$

$$= \left[\frac{e_{21}}{\mu_{21}} + \frac{e_{22}}{\mu_{22}} \right]^{k_2} ,$$

where again the last expression is obtained from the multinomial theorem. Multiplying and dividing this expression by $e_{21}+e_{22}$ gives

$$= \quad (\bar{e}_{21} \pm e_{22}) \quad \left[\frac{e_{21}}{e_{21}+e_{22}} \; \frac{1}{\mu_{21}} + \frac{e_{22}}{e_{21}+e_{22}} \; \frac{1}{\mu_{22}} \right]^{k_2} .$$

Since the e's are relative throughputs the term inside the brackets is the average service time and $e_{21}+e_{22}$ the relative throughput of a chain j customer. That is,

$$w_2(z_2) = \left[\frac{e_{2j}}{\mu_{2j}} \right]^{k_j} ,$$

where $k_j = k_2$, $e_{2j} = e_{21}+e_{22}$, and

$$\frac{1}{\mu_{2j}} = \frac{e_{21}}{e_{21}+e_{22}} \; \frac{1}{\mu_{21}} + \frac{e_{22}}{e_{21}+e_{22}} \; \frac{1}{\mu_{22}} .$$

Now since there were only two classes at service center 2 and both of these belonged to chain j, it should be obvious that $h_2(k_2) = w_2(k_2)$. In fact the expression for $h_2(k_2)$ was derived earlier as an intermediate step and is:

$$h_2(k_2) = \left[\frac{e_{21}}{\mu_{21}} + \frac{e_{22}}{\mu_{22}} \right]^{k_2} .$$

6.5.4 Open Networks with a Constant Mean Arrival Rate

For an open network with a constant mean arrival rate, it is possible to obtain closed form solutions for the normalizing constant and steady-state probabilities. If mean arrival rate is constant, then the marginal steady-state probabilities are given by

$$P(k_1,k_2,\ldots,k_N) = \lambda^{K-1} \frac{h_1(k_1)\ h_2(k_2)\ \ldots\ h_N(k_N)}{G} , \qquad (6.22)$$

where $K = k_1 + k_2 + \cdots + k_N$. Multiplying the numerator and denominator by λ and letting G absorb the λ in the denominator results in:

$$P(k_1,k_2,\ldots,k_N) = \frac{\{\lambda^{k_1}\ h_1(k_1)\}\ \{\lambda^{k_2}\ h_2(k_2)\}\ \cdots\ \{\lambda^{k_N}\ h_N(k_N)\}}{G} . \qquad (6.23)$$

The normalizing constant is determined by summing this expression over all feasible states. That is,

$$G = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \cdots \sum_{k_N=0}^{\infty} \left[\{\lambda^{k_1}\ h_1(k_1)\}\ \{\lambda^{k_2}\ h_2(k_2)\}\ \ldots\ \{\lambda^{k_N}\ h_N(k_N)\} \right]$$

$$= \left[\sum_{k_1=0}^{\infty} \lambda^{k_1}\ h_1(k_1) \right] \left[\sum_{k_2=0}^{\infty} \lambda^{k_2}\ h_2(k_2) \right] \cdots \left[\sum_{k_N=0}^{\infty} \lambda^{k_N}\ h_N(k_N) \right] . \qquad (6.24)$$

Hence, the expression for G factors into terms where each term involves only parameters for a single service center. That is,

$$G = \mathop{}_{n=1}^{N} G_i ,$$

where

$$G_i = \sum_{k_i=0}^{\infty} \lambda^{k_i}\ h_i(k_i) . \qquad (6.25)$$

It follows from the definitions of G_i and $P_i(0)$ that:

$$P_i(0) = \frac{1}{G_i} .$$

It also follows from the geometric series and the series expansion of $e^{x}(\exp^{x})$, that

$$P_i(0) = \begin{cases} 1 - \sum\limits_c \lambda(e_{ic}/\mu_{ic}) & \text{for i FCFS/1/, PS or LCFSPR} \\ \exp\left[\sum\limits_c - \lambda(e_{ic}/\mu_{ic})\right] & \text{for i IS .} \end{cases} \qquad (6.26)$$

Let

$$\rho_i = \begin{cases} \sum\limits_c \lambda(e_{ic}/\mu_{i_c}) & \text{if i FCFS, PS or LCFSPR} \\ \sum\limits_c \lambda(e_{ic}/\mu_{ic}) & \text{if i IS .} \end{cases} \qquad 6.27)$$

Note that ρ_i is the utilization of type FCFS/1/, PS, and LCFSPR service centers since $\rho_i = 1 - P_i(0)$ for all single server, service centers. However, the utilization of type IS service centers is by definition zero. Returning to the problem at hand, it follows that the number of customers in each service center is an independent random variable. More precisely,

$$P(k_1,k_2,\ldots,k_N) = P_1(k_1)\ P_2(k_2)\ \ldots\ P_N(k_N)$$

where

$$P_i(k_i) = \begin{cases} \rho_i^{k_i}\ (1-\rho_i) & \text{if i FCFS/1/, PS, or LCFSPR} \\ (\rho_i^{k_i}/k_i!)\ e^{-\rho_i} & \text{if i IS.} \end{cases} \qquad (6.28)$$

The results are amazing! For FCFS/1/, PS, and LCFSPR service centers the marginal distribution is the same as the distribution of customers in an M/M/1 system. If the service center has an infinite number of servers, then the marginal distribution is the same as the distribution of an M/M/∞ (with an appropriately defined ρ_i). Moreover, the network steady-state probabilities factor into products with one term for each service center, and each term is the solution for that service center 'in isolation' with a Poisson input and with an exponential service time and appropriately defined ρ_i.

It follows that since the distribution of customers are the same that the mean values are the same. That is,

$$L_i = \begin{cases} \dfrac{\rho_i}{1 - \rho_i} & \text{if } i \text{ FCFS/1/, PS, or LCFSPR} \\ \rho_i & \text{if } i \text{ IS.} \end{cases} \tag{6.29}$$

Since the network is open, the mean rate that customers enter the network equals the mean rate that customers leave the network. Thus, the throughput equals the mean rate that customers enter the network. The same is true at the service center level. Since e_{ic} equals the mean number of visits a customer makes to service center i, class c, it follows that the throughput of service center i is

$$T_i = \lambda \sum_c e_{ic} \quad . \tag{6.30}$$

Given the mean number of customers and the throughput, the mean response time can be calculated from Little's law. Usually it is the performance measure of the service centers, not the overall network, that is of interest.

If there is more than one service center, the distribution of the response time cannot be determined. This is because the response times at the service centers are not independent [BURK64]. Notice that if a network consists of a single service center with all customer classes having the same exponential service time, then the distribution of customers, throughput, and mean response time are the same for service center types FCFS/1/, PS, and LCFSPR. As pointed out by Kleinrock, although the mean response times are the same, there is a large difference in the variances. One may, therefore, conclude that the average response time by itself is not a very good indicator of system performance [KLEI76].

6.5.6 Load Dependent Service Rates and Multiple Servers

The service rate at a service center is said to be load dependent if the rate that customers are served depends upon the number of customers at the service center. For example, if a service center contains identical and multiple servers then the service rate is a function of both the number of customers and the number of servers. In the previous chapters this has been expressed as:

$$\mu_i(k_i) = \begin{cases} k_i \, \mu_i & \text{for } k_i \leq m_i \\ m_i \, \mu_i & \text{for } k_i \geq m_i \end{cases}$$

where

k_i = the number of customers at service center i

m_i = the number of identical servers

μ_i = the service rate when $k_i = 1$. (6.31)

The class of networks discussed in this chapter can be extended not only to include this case, but more general cases as well. If $Z_i(k_i)$ is the relative service rate (relative to μ_i which is the service rate when k_i=1) at service center i when there are k_i customers at service center i then

$$f_i(x_i) := f_i(x_i) \prod_{a=1}^{k_i} 1/Z_i(a) \qquad (6.32)$$

$$g_i(y_i) := g_i(y_i) \prod_{a=1}^{k_i} 1/Z_i(a) \qquad (6.33)$$

$$w_i(z_i) := w_i(z_i) \prod_{a=1}^{k_i} 1/Z_i(a) \qquad (6.34)$$

$$h_i(k_i) := h_i(k_i) \prod_{a=1}^{k_i} 1/Z_i(a) \qquad (6.35)$$

where := is an assignment operator and is read as 'becomes'. The proof of these equations is identical to the one in the previous chapter section 5.4.5 where the subset of service centers is simply service center i. The only restrictions on $Z_i(k_i)$ are that it be a positive function of k_i and $Z_i(1)$=1. However, it is usually assumed that there exist some m such that for all $k_i \geq m$ $Z_i(k_i)=Z_i(m)$. For example, for the identical and multiple service center case

$$Z_i(k_i) = \begin{cases} k_i & \text{for } k_i \leq m_i \\ m_i & \text{for } k_i \geq m_i \,, \end{cases} \qquad (6.36)$$

and the service rate is μ_i $Z_i(k_i)$. Such service centers are often referred to as 'limited load dependent' centers because the service rate depends upon k_i only up to m. It will be assumed throughout this chapter and the next that all load dependent service centers are of this type.

Now let
$$\xi_i = \sum_{c} \lambda(e_{ic}/\mu_{ic}) \quad , \tag{6.37}$$

and $\rho_i = \xi_i/Z_i(m)$. Also let G_i equal the component of the normalizing constant that is due to service center i. That is $G_i = 1/P_i(0)$. It follows that

$$G_i = \sum_{k_i=0}^{\infty} \lambda_i^{k_i} h_i(k_i) = \sum_{k_i=0}^{\infty} \xi_i^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a)$$

$$= \sum_{k_i=0}^{m-1} \xi_i^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a) + \sum_{k_i=m}^{\infty} \xi_i^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a)$$

$$= \sum_{k_i=0}^{m-1} \xi_i^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a) + \xi_i^{m} \prod_{a=1}^{m} 1/Z_i(a) \sum_{k_i=0}^{\infty} [\xi_i/Z_i(m)]^{k_i} \quad .$$

$$= \frac{\xi_i^{m}}{\prod_{a=1}^{m} Z_i(a)\ (1-\rho_i)} + \sum_{k_i=0}^{m-1} \frac{\xi_i^{k_i}}{\prod_{a=1}^{k_i} Z_i(a)} \quad . \tag{6.38}$$

Similarly, the mean customer population at service center i is

$$L_i = \frac{1}{G_i} \sum_{k_i=1}^{\infty} k_i\, \lambda^{k_i} h_i(k_i) = \frac{1}{G_i} \sum_{k_i=1}^{\infty} k_i\, \xi_i^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a)$$

$$= \frac{1}{G_i} \sum_{k_i=1}^{m-1} k_i\, \xi_i{}^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a) + \frac{1}{G_i} \sum_{k_i=m}^{\infty} k_i\, \xi_i{}^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a)$$

$$= \frac{1}{G_i} \sum_{k_i=1}^{m-1} k_i\, \xi_i{}^{k_i} \prod_{a=1}^{k_i} 1/Z_i(a) + \frac{\xi_i{}^m}{G_i} \prod_{a=1}^{m} 1/Z_i(a) \sum_{k_i=0}^{\infty} (k_i+m)\, \rho_i{}^{k_i}$$

$$= \frac{k_i\, \xi_i{}^m}{\prod_{a=1}^{m} Z_i(a)} \left[\frac{m}{(1-\rho_i)} + \frac{\rho_i}{(1-\rho_i)^2} \right] + \frac{1}{G_i} \sum_{k=1}^{m-1} \frac{k_i\, \xi_i{}^{ik}}{\prod_{a=1}^{k_i} Z_i(a)} \quad . \quad (6.39)$$

For the the special case of

$$Z_i(k_i) = \begin{cases} k_i & \text{for } k_i \leq m_i \\ m_i & \text{for } k_i \geq m_i \, , \end{cases} \qquad (6.40)$$

G_i and L_i reduce to

$$G_i = \sum_{k_i=1}^{m_i-1} \frac{(m_i\rho_i)^{k_i}}{k_i!} + \frac{(m_i\rho_i)^{m_i}}{m!\,(1-\rho_i)} \qquad (6.41)$$

$$L_i = m_i\rho_i + \frac{\rho_i\,(m_i\rho_i)^{m_i}}{G_i\, m!\,(1-\rho_i)^2} \; . \qquad (6.42)$$

It should not be surprising that G_i is identical to $1/P_0$ for the M/M/m system and L_i equals L of the M/M/m system. Also note that ρ_i is in agreement with the utilization of the M/M/m system.

There are, however, cases when one might want to assign $Z_i(k_i)$

differently. In these cases the more complicated expressions must be used. In closing, the results of this section are valid for FCFS, PS, and LCFSPR service disciplines. However, unless it is explicitly stated that the service rate is load dependent, it will be assumed to be load independent. Furthermore, the notation FCFS will be assumed to imply FCFS/1/ (load independent) unless stated otherwise.

6.5.7 An Example of Queueing Networks with Multiple Classes

The following is the only known example of an open queueing network with more than one class of customers and more than one service center. The example is from 'Computer Performance Modeling Handbook' [LAVE83]. The example was not intended to illustrate the analysis of queueing networks, but to show that the number of classes required to describe the routing can be significantly reduced by combining sources. Unfortunately, the example as it appears in the reference does not contain a figure showing the routing, and therefore is very difficult to follow. In addition, all equations and results are stated in sentence form. The example as it appears here has been greatly expanded.

Figure 6.15 depicts the model of a small communications network organized as a ring. Messages (customers) originate at the sources and terminate at the sinks. Every message must pass through one or more communication links (service centers). For example, a message from source 1 goes through link 1 to get to destination 2, through links 1 and 2 to get to destination 3, and through links 1,2, and 3 to get to destination 4. A message from source 2 goes through link 2 to get to destination_3, through links 2 and 3 to get destination 4, and through links 2,3, and 4 to get to destination 1. Messages from sources 3 and 4 behave similarly.

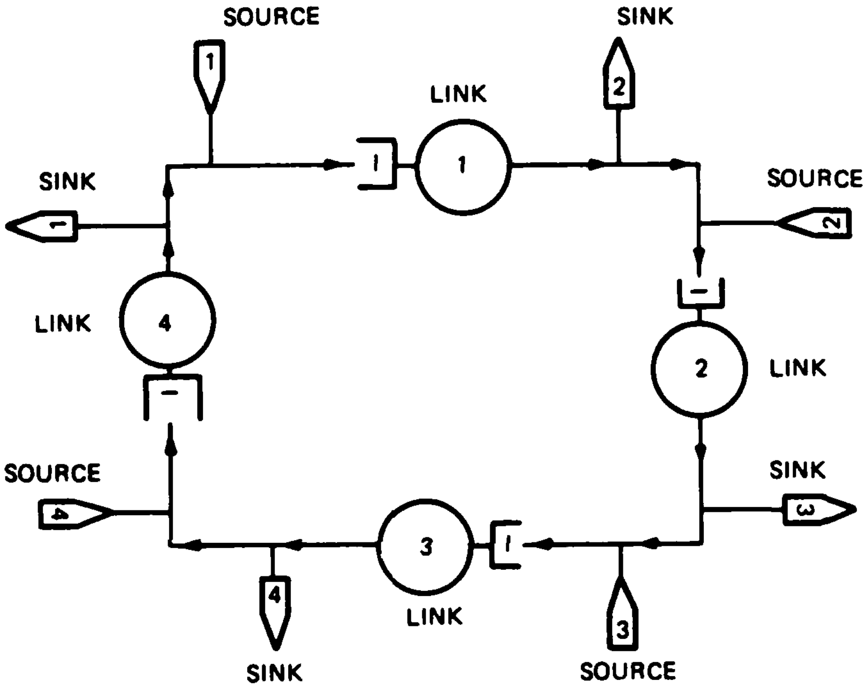


Figure 6.15 Model of a Communications System.

The routing of messages in the network cannot be described by a single set of routing probabilities. This is because the routing probabilities depend on where the message originated. For example a message from source 1 that passes through link 3 must terminate at destination 4, whereas messages from the other sources may be routed through link 4. In order to describe the routing, it is necessary to partition the messages into classes. If it is assumed that there are four sources, then six classes per service center are required to describe the routing (see Figure 6.16). For example, source-destination pairs of messages that pass through link two are : (2,3),(2,4),(2,1),(1,3),(1,4), and (4,3).

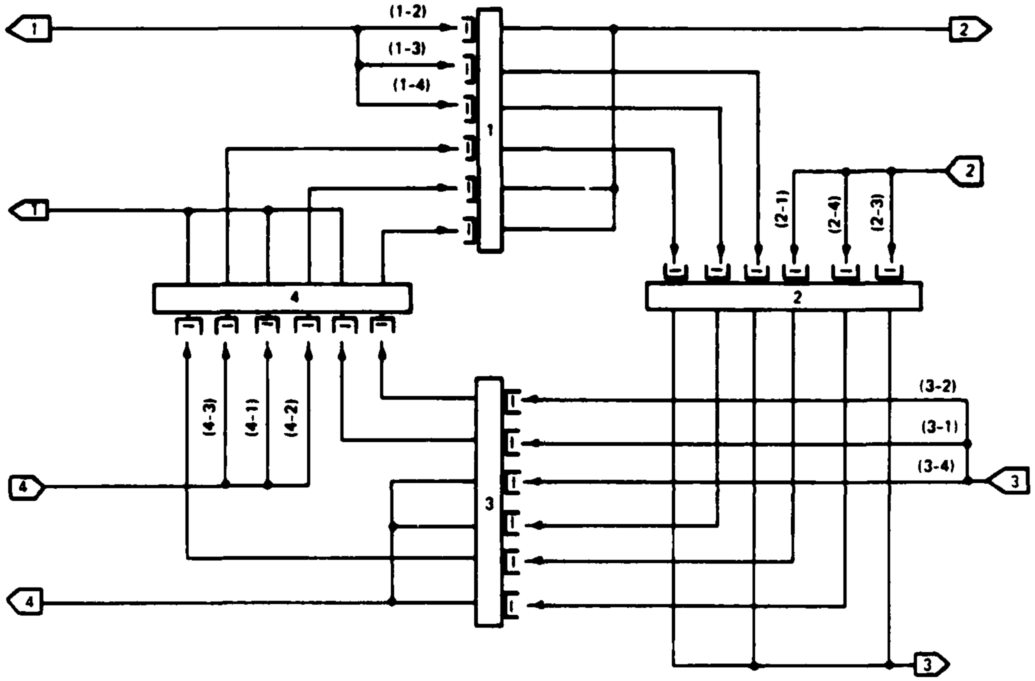


Figure 6.18 Classes Required for Routing with Multiple Sources.

On the other hand the number of classes at each service center can be reduced by combining the four sources into a single (aggregate) source. This reduces the number of classes at each service center required to describe the routing to three, one for each possible destination (see Figure 6.17). For example at service center 2 classes are required for destinations 3,4 and 1. Thus, a minimum of 12 classes are required in order to describe the routing. In order to obtain a solution, it is necessary that all messages have the same service time distribution, and that it be an exponential.

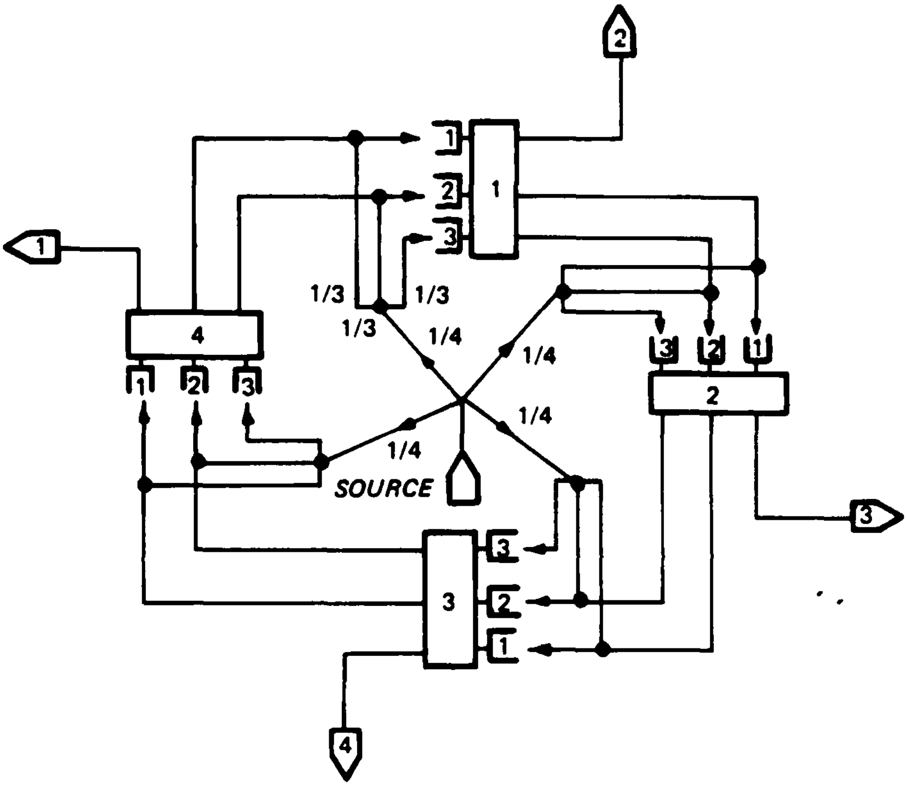


Figure 6.17 Classes Required with an Aggregate Source.

It is assumed that messages arrive from the four sources at a rate of 2.5 messages per second and all destinations are equally likely. It is also assumed that the mean message length is 360 bits and the transmission rate is 2400 baud. Thus, the arrival rate (λ) of messages from the composite source is 10 messages per second, 1/4 of these messages are routed to each service center, or 1/12 to each class at each service center. The service rate (μ_i) is 2400/360 or 6.67 messages per second (the message length includes overhead). The relative throughput equations are :

$$e_{11} = 1/12 + e_{42}$$

$$e_{12} = 1/12 + e_{43}$$

$$e_{12} = 1/12$$

$$e_{21} = 1/12 + e_{12}$$

$$e_{22} = 1/12 + e_{13}$$

$$e_{22} = 1/12$$

$$e_{31} = 1/12 + e_{22}$$

$$e_{32} = 1/12 + e_{23}$$

$$e_{32} = 1/12$$

$$e_{41} = 1/12 + e_{32}$$

$$e_{42} = 1/12 + e_{33}$$

$$e_{42} = 1/12 .$$

Solving these equations results in :

$$e_{11} = e_{21} = e_{31} = e_{41} = 1/4$$

$$e_{12} = e_{22} = e_{32} = e_{42} = 1/6$$

$$e_{13} = e_{23} = e_{33} = e_{43} = 1/12 .$$

The utilization at each service center is :

$$\rho_i = \sum_c \lambda(e_{ic}/\mu_i) = (\lambda/\mu_i)\ (e_{i1}+e_{i2}+e_{i3})$$

$$= (10/6.67)(1/4+1/6+1/12) = 0.75 \quad .$$

The distribution of messages at each service center is:

$$P_i(k_i) = (0.75)^{k_i}\,(1-0.75) \quad .$$

The mean number of messages at each service center is

$$L_i = \frac{\rho_i}{1-\rho_i} = 0.75/(1-0.75) = 3 \text{ messages.}$$

The throughput at each service center is

$$T_i = \lambda \sum_c e_{ic} = 10\ (1/4+1/6+1/12) = 5 \text{ messages per second.}$$

The mean response time at each service center is

$$R_i = \frac{L_i}{T_i} = 3/5 = 0.6 \text{ seconds per messages.}$$

Thus, the mean response times of messages that pass through 1,2 and 3 service centers are 0.6s, 1.2s and 1.8s respectfully. The mean response time of an average message is the weighted sum of the means. Since all means are equally likely, the mean response time of an average message is: (0.6)(1/3)+(1.2)(1/3)+(1.8)(1/3) = 1.2 seconds.

One of the problems facing the designer of a communication network is buffer size. If it is too small then messages will be lost. In order to properly determine the buffer size, one needs to know the distribution of messages at each service center. That is, more information is required than just the mean. For example, the probability that the service centers in Figure 6.8 contains more than 6 customers is:

$$P(k_i > 6) = 1 - [P(0)+P(1)+(2)+(3)+(4)+P(5)+P(6)] = 0.1334 \ .$$

That is, 13.34 percent of the time a service center contains more that 6 customers. Hence, even though the mean is only 3 messages a buffer size of 6 is inadequate!

CHAPTER 7

CLOSED AND MIXED NETWORKS

7.1 Introduction

The equations derived in the last chapter give only the form of the solution for closed and mixed networks. More specifically, a closed form expression for the normalizing constant, G, was not, and in general cannot be obtained. The reason for this is that the number of customers in a closed chain remains constant, and therefore the distribution of customers at the individual service centers are not independent. By definition, the normalizing constant, G, is the summation of all unnormalized probability distributions over all feasible networks states. Clearly, for a closed network the number of feasible networks states equals the number of possible ways that customers can be distributed over the network, such that the number of customers in each chain remains constant. Unfortunately this number increases rapidly as the number of service centers and/or the number of customers increases. The reader is referred to the example in Chapter 5 in which a network consisting of a single chain with eight service centers and 20 customers yielded 888,030 network states. It should also be noted that each term in the summation contains virtually every parameter associated with every service center in the network. Thus, even for the smallest of networks the expression for G would probably be too complex to determine how the parameters affect performance. Fortunately, there are other ways of determining G than directly from its definition. However, any expression derived from these is still so complex that usually all one can handle is the numerical result.

At the present there are three algorithms for determining the performance metrics of closed networks : the Convolution algorithm, the Mean Value Analysis (MVA) algorithm, and the Local Balance Algorithm for Normalizing Constants (LBANC). All extend to the full class of networks known to have a product form solution. However, no one of them is best for solving all problems on all machines.

Before discussing the advantages and disadvantages of these algorithms, it is first best to make some general comments and define some notations. All of the algorithms have recursive equations. That is, either the equation for the normalizing constant or some other parameter is given in terms of the same parameter with one less customer in the system. More precisely, one less customer in one of the chains. For a network with J closed chains let $V = (V_1,V_2,...,V_J)$ where V_j is the number of customers in the jth chain. The vector V will be referred to as the population vector. Now let 1_j be a vector with a one in the jth component and zeros everywhere else. Thus, the vector $V-1_j$ represents the network with one less customer in chain j.

The convolution algorithm was developed by Buzen in the early 1970's [BUZE73]. Its name comes from fact that the recurrence equation for G(V) resembles discrete convolution. In the case of a single chain closed network, G(0) is assigned the value of one. This value is used to compute G(1), the normalizing constant when there is one customer in the system. The value of G(1) is then used to compute G(2) and so on.

The process is repeated until the desired population has been reached. For a single chain network with N service centers, the algorithm requires approximately $N+V$ multiplications and N+V additions. Unfortunately, the algorithm is sensitive to the value that must be assigned to one of e's, and numerical problems can occur regardless of what value is assigned (usually overflow). In addition, the complexity of the algorithm increases significantly for networks with multiple chains. The details of the algorithm will not be discussed here due to the lack of room and the fact it is discussed elsewhere. The interested reader is referred to 'Computational Algorithms for Closed Queueing Networks' by Bruell [BRUE80]. This reference is a condensed version of his Ph.D dissertation. It contains over 200 pages and starts with the equations developed in the previous chapter . Approximately two-thirds of the book is devoted to this one algorithm.

In the late 1970's Reiser and Lavenberg developed an iterative algorithm that can determine mean performance values without finding the normalizing constant [REIS80]. Hence, the name Mean Value Analysis. For single chain networks the mean values are determined for the network with one customer. These values are used to determine the mean values of the network with two customers, and so on. The algorithm requires approximately the same number of computations and storage as the convolution algorithm, however it is less sensitive to numerical problems. It is the algorithm of choice except for networks that contain several service centers with 'limited load dependent' service rates (See Chapter 6, Section 6.5.6) [BRUE80] [LAVE83] [HAYE84]. In addition it is the only algorithm whose equations have intuitive meanings. This not only makes it easier to explain, but more importantly, easier to remember. For these reasons and because there exist no examples in the literature, it will be discussed here. However, it will not be discussed with the same mathematical rigor as the previous chapters. There are several reasons for this. The first is that many of the proofs employ results that are from the convolution algorithm, which would have to be explained and derived also. Another reason for not deriving all of the equations is that this would conflict with presenting the material in a tutorial fashion, which is a primary objective. That is, many of the simplest to use equations are special cases of more general equations, and they cannot be derived without first deriving the more general ones.

LBANC was inspired by MVA and closely parallels it [CHAN80]. In addition to determining the mean performance values, the normalizing constant is also determined as implied in its name. It is the algorithm of choice when probability distributions are required. However, it has the same numerical stability problems as the convolution algorithm, and the same disadvantages as MVA when dealing with limited load dependent service rates. Another disadvantage is that its equations do not have intuitive meaning and are difficult to remember. With the exception of a few comments, LBANC will not be discussed further.

7.2 Closed Networks

For a closed network, the complexity of determining the performance metrics is directly proportional to the number of feasible network states, and as stated in the last chapter the number of feasible network states can significantly be reduced by eliminating classes and distinguishing between customers according to chains. Although this is the starting point of all algorithms for closed networks, the authors of these algorithms state this in very obscure ways. For example Reiser, the author of the MVA algorithm, simply states that a network that allows customers to switch classes can be mapped into a model without class changes [REIS80]. Although this statement might be implied in the reference he gives, there is no such statement in the paper. Bruell starts transforming the classes into what he called equivalent classes without giving a reference or justification for doing so [BRUE80]. In addition, several of Bruell's statements and equations concerning obtaining class metrics from equivalent class metrics are wrong. Chandy and Sauer, the authors of LBANC, make the statement that it is more convenient to first obtain metrics by service centers and then obtain class metrics from these (their statement as given is valid only for single chain networks) [CHAN80]. Again, they give no reference or justification. In addition, many of the authors that have made extensions or modifications to the original algorithms make statements such as each chain j customer belongs to the same customer class, or that the term class and chain are used synonymously. They make no mention of the fact that the original algorithms or their modified ones can handle networks in which customers are allowed to change classes. It is believed that at least part of the confusion arises from the fact that there is no reference explicitly explaining this. More precisely, it is believed that the aggregate state in which customers are distinguished according to chains, along with its justification given in the previous chapter, is not new but does not appear (at least explicitly) elsewhere in the literature.

Considering the problem at hand, once the chain performance metrics have been determined, class performance metrics can be easily calculated. For example, once the normalizing constant has been determined its value can simply be substituted into the equation for probability distribution by class. Before deriving the equations for converting chain performance metrics to chain metrics, a notational change will be introduced. The mean service rates of a class c or chain j customer, denoted μ_{ic} and μ_{ij} respectively, more often than not have appeared in their reciprocal form. Therefore, let

$$1/\mu_{ic} = s_{ic} \tag{7.1}$$

$$1/\mu_{ij} = s_{ij} \ . \tag{7.2}$$

Obviously, s_{ic} and s_{ij} are the mean service times of a class c and chain j customer at service center i. The equations for determining mean class metrics from mean chain metrics and their derivations follow:

Throughput by class can be determined from the known relative throughputs e_{ic} and e_{ij}. More precisely,

$$T_{ic}(V) = (e_{ic}/e_{ij})\ T_{ij}(V) \ . \tag{7.3}$$

Utilization by class can be determined from the equation :

$$\rho_{ic}(V) = [(e_{ic}s_{ic})/(e_{ij}s_{ij})]\ \rho_{ij}(V) \ , \tag{7.4}$$

which follows from the reasoning that since the e's are relative throughputs, $e_{ic}s_{ic}$ and $e_{ij}s_{ij}$ are relative utilizations. Similarly, the mean number of class c customers at service center i is given by :

$$L_{ic}(V) = [(e_{ic}s_{ic})/(e_{ij}s_{ij})]\ L_{ij}(V)\ , \qquad (7.5)$$

which follows from the fact that the term in brackets (the ratio of relative utilizations) is the conditional probability that an arbitrary customer, waiting for service or already receiving service, is in class c, given that it is in chain j. Class response time can be calculated from Little's law. More precisely,

$$R_{ic}(V) = L_{ic}(V)/T_{ic}(V)\ . \qquad (7.6)$$

7.2.1 <u>The Arrival Theorem</u>

Mean performance parameters for a queueing network with multiple closed chains and a product form solution can be determined from the three principles :

(1) A chain j customer arriving at service center i 'sees' the system with himself removed and in equilibrium,

(2) Little's Law applies to chains,

(3) Little's Law applies to service centers.

The first of the these principles is known as the arrival theorem. It's proof, [LAVE79], depends on results that can only be derived from the convolution algorithm and will not be repeated here. It is important to emphasize that the arrival theorem only holds for networks that have a product-form solution. Most of this chapter is concerned with the application of these three principles.

7.2.2 The Throughput Theorem

As its name implies, MVA deals with determining the mean values of performance metrics such as throughput, response time, and customer distribution. However, it is possible to determine the normalizing constant from these values via the throughput theorem. It states that the average throughput of a chain j customer through service center i is :

$$T_{ij}(V) = \frac{G(V-1_j)}{G(V)} \; e_{ij} \, , \qquad (7.7)$$

where $G(V-1_j)$ is the normalizing constant of the network with one less customer in chain j. The throughput theorem is one of the primary results of the convolution algorithm and will not be proven here. Solving this equation for G(V) results in

$$G(V) = \frac{G(V-1_j)}{T_{ij}(V)} \; e_{ij} \, . \qquad (7.8)$$

Now, since e_{ij} is known and MVA requires the calculation of $T_{ij}(V)$ for all V up to the desired population, G(V) can easily be determined. The procedure will be illustrated later by examples. Unfortunately, this step does add to the storage requirement of the algorithm. This may or may not be an issue depending on the size and population of the network and the amount of usable memory.

At this point some additional comments about the normalizing constant are appropriate. As of the present there is no algorithm or scaling technique that will always prevent overflow from occurring when trying to calculate the normalizing constant. One of the primary advantages of MVA over convolution and LBANC is that it doesn't require the normalizing constant to obtain performance metrics.

7.2.3 Single Chain - Load Independent - Closed Networks

In order to illustrate the MVA algorithm consider a single chain closed network with N, load independent, single server, FCFS, service centers. Clearly, the mean time a customer stays at a service center is his mean service time plus the mean time it takes for the service center to dispose of the backlog of customer ahead of it. Since the mean service time of all customers at service center i is the same, the mean response time of a customer at service center i is:

$$R_i(V) = s_i \, [1 + L_i(V-1)], \tag{7.9}$$

where the term $L_i(V-1)$ is the average backlog of customers and follows directly from the arrival theorem.

Throughput can now be calculated from response times by the equation:

$$T_i(V) = V \Big/ \sum_{n=1}^{N} (e_n/e_i) \, R_n(V) \ . \tag{7.10}$$

The equation for throughput follows from the fact that (e_n/e_i) is the number of times a customer visits service center n before returning to service center i, and therefore the summation is the time it takes a customer to pass through service center i and return.

The mean number of customers at each service center can now be calculated from Little's law. More precisely :

$$L_i(V) = T_i(V)\ R_i(V)\ . \tag{7.11}$$

Thus, $R_i(0)$, $\bar{T}_i(0)$, and $L_i(0)$ can be calculated and these values used to determine $R_i(1)$, $T_i(1)$, $L_i(1)$, the results used to determine the performance metrics for V=2 and so on. At any point in the calculations, the utilization of the service center can be calculated from

$$\rho_i(V) = s_i\ T_i(V). \tag{7.12}$$

Although it was assumed that all service centers were load independent FCFS, the equations are valid also for load independent PS and LCFSPR service centers. This follows from the fact that the aggregate state probabilities are the same, and consequently, so are the mean performance metrics. The infinite server case is trivial. That is, since the number of servers is always greater than or equal to the number of customers, the mean response time is just the mean service time. A more program-like definition of the algorithm is given in Table 7.1. Notice that the throughputs for all but one of service centers are obtained from their relative throughputs.

As an example, the closed network of Figure 5.7 will be reworked using MVA. For convenience, service rates and relative throughputs are repeated here:

s_1 = 10ms	e_1 = 100
s_2 = 25ms	e_2 = 80
s_3 = 100ms	e_3 = 10 .

```
Begin
For i:=1 to N do {initialization}
  $L_i(0) := 0$

 For v=0 to V do {body}
 begin

 {response time}
   For i:=1 to N do  {response time}

      $R_i(V) = \begin{cases} s_i\ [1 + L_i(V-1)] & \text{if i FCFS, PS, LCFSPR} \\ s_i & \text{if i IS} \end{cases}$

 {throughput}

   $T_1(v) = e_1\ v\ /\ \sum_{j=1}^{N} e_j\ R_j(v)$

   For i:=2 to N do

   $T_i(v) = (e_i/e_1)\ T_1(v)$

 {queue length}
   For i:=1 to N do

      $L_i(V) = T_i(V)\ R_i(V)$

 end {MVA body}

{utilization}
 For i:=1 to N do

    $\rho_i(V) = \begin{cases} s_i\ T_i(V) & \text{if i FCFS, PS, or LCFSPR} \\ 0 & \text{if i IS} \end{cases}$

End.
```

Table 7.1 MVA Algorithm For Single Chain, Load Independent, Networks.

Calculations for $v = 1$:

$$R_1(1) = (10*10^{-3})\ (1)$$

$$R_2(1) = (25*10^{-3})\ (1)$$

$$R_3(1) = (100*10^{-3})\ (1)$$

$$T_1(1) = \frac{(1)\ (100)}{(100)(10*10^{-3}) + (80)(25*10^{-3}) + (10)(100*10^{-3})} = 25$$

$$T_2(1) = (80/100)\ T_1(1) = 20$$

$$T_3(1) = (10/100)\ T_1(1) = 2.5$$

$$L_1(1) = (25)\ (10*10^{-3}) = 0.25$$

$$L_2(1) = (20)\ (25*10^{-3}) = 0.50$$

$$L_3(1) = (2.5)\ (100*10^{-3}) = 0.25\ .$$

Calculations for $v = 2$:

$$R_1(2) = (10*10^{-3})\ (1.25) = 12.5*10^{-3}$$

$$R_2(2) = (25*10^{-3})\ (1.5) = 37.5*10^{-3}$$

$$R_3(2) = (100*10^{-3})\ (1.25) = 125*10^{-3}$$

$$T_1(2) = \frac{(2)\ (100)}{(100)(12.5*10^{-3}) + (80)(37.5*10^{-3}) + (10)(125*10^{-3})} = 36.36$$

$$T_2(2) = (80/100)\ T_1(2) = 29.09$$

$$T_3(2) = (10/100)\ T_1(2) = 3.636$$

$$L_1(2) = (36.36)\ (12.5*10^{-3}) = 0.445$$

$$L_2(2) = (29.09)\ (37.5*10^{-3}) = 1.091$$

$$L_3(2) = (3.636)\ (125*10^{-3}) = 0.445\ .$$

Calculations for $v = 3$:

$R_1(3) = (10*10^{-3})\ (1.455) = 14.55*10^{-3}$

$R_2(3) = (25*10^{-3})\ (2.091) = 52.28*10^{-3}$

$R_3(3) = (100*10^{-3})\ (1.445) = 145.5*10^{-3}$

$$T_1(2) = \frac{(3)\ (100)}{(100)(14.55*10^{-3}) + (80)(52.28*10^{-3}) + (10)(145.5*10^{-3})}$$

$$= 42.3$$

$T_2(2) = (80/100)\ T_1(3) = 33.84$

$T_3(2) = (10/100)\ T_1(3) = 4.23$

$L_1(3) = (42.3)\ (14.55*10^{-3}) = 0.6155$

$L_2(3) = (33.84)\ (52.28*10^{-3}) = 1.769$

$L_3(3) = (4.23)\ (145.5*10^{-3}) = 0.6155$

$\rho_1(3) = (10*10^{-3})\ (42.3) = 0.423$

$\rho_2(3) = (25*10^{-3})\ (33.84) = 0.846$

$\rho_3(3) = (100*10^{-3})\ (4.23) = 0.423$.

If desired the normalizing constants can now be calculated from the throughputs. More precisely,

$$G(V) = \frac{G(V-1)}{T_1(V)}\ e_1 ,$$

$$G(0) = 1 ,$$

$$G(1) = \frac{(1)(100)}{25} = 4 ,$$

$$G(2) = \frac{(4)(100)}{36.36} = 11 ,$$

$$G(3) = \frac{(11)(100)}{42.3} = 26 .$$

7.2.4 Single Chain - Load Dependent - Closed Networks

Let $s_i(k_i)$ equal $1/\mu_i(k_i)$. For service centers with limited load dependent service rates the average response is obtained from:

$$R_i(V) = \sum_{k_i=1}^{V} k_i \; s_i(k_i) \; P_i(k_i-1|V-1) , \qquad (7.13)$$

where $P_i(k_i-1|V-1)$ is the marginal probability of finding k_i-1 customers at service center i, given that the network contains V-1 customers. It can be determined from the recurrence equation:

$$P_i(k_i|V) = \begin{cases} 1 & \text{for } k_i=0 \text{ and } V=0 \\ s_i(k_i)\; T_i(V)\; P_i(k_i-1|V-1) & \text{for } k_i>0 \\ 1 - \sum_{k_i=0}^{V} P_i(k_i|V) & \text{for } k_i=0 \text{ and } V>0 . \end{cases} \qquad (7.14)$$

Note that the equation for the response time holds even if the service rate is fixed or if it is strictly load dependent (infinite servers). Although it will not be proven here, the equation for the marginal probability also holds. The reason for this is that both equations are intermediate steps in the proof of the arrival theorem. More precisely, the equations for multiple chain load dependent service centers (the most complex case) are derived first and all the others from these. The utilization then of a limited load dependent service center is :

$$\rho_i(V) = \sum_{k_i=0}^{V} \min(k_i,m_i)\ P_i(k_i|V)\ /\ m_i\ . \qquad (7.15)$$

This follows from the fact that when service center i contains k_i customers, $[\min(k_i,m_i)]/m_i$ is the capacity of the service center that is being used.

The procedure will be illustrated by the network in Figure 7.1. A description of the service centers is given in Table 7.2. It follows from the routing probabilities in Figure 7.1 that if e_2 is assigned the value of one, then $e_1 = e_3 = 0.5$. Thus,

$$T_2(V) = (e_2/e_1)\ T_1(V) = (2)\ T_1(V) \text{ for all } V$$

$$T_3(V) = (e_3/e_1)\ T_1(V) = (1)\ T_1(V) \text{ for all } V.$$

It follows from the service rates of the individual servers that:

$$s_1 = 2$$

$$s_2 = 1$$

$$s_3(k_i) = \begin{cases} 4 \text{ for } k_i=1 \\ 2 \text{ for } k_i>1\ . \end{cases}$$

The following are the calculations for v=1, v=2, and v=3:

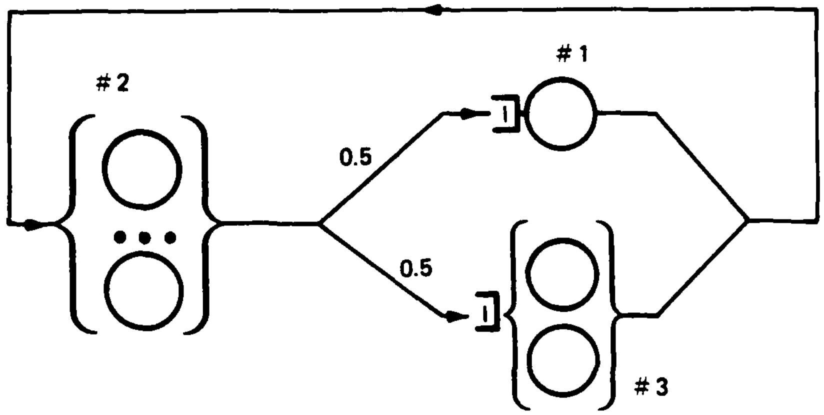


Figure 7.1 Example of a Load Dependent Network.

Service Center	Number of Servers	Service Discipline	Server Rate μ_i
1	1	PS	0.5
2	∞	IS	1
3	2	FCFS	0.25

Table 7.2 Description of Service Centers in Figure 7.1.

Calculations for $v=1$:

$R_1(1) = (2)(1) = 2$

$R_2(1) = 1$

$R_3(1) = (4)\ P_3(0|0) = (4)\ (1) = 4$

$$T_1(1) = \frac{(0.5)\ (1)}{(0.5)(2) + (1)(1) + (0.5)(4)} = 0.125$$

$T_2(1) = (2)\ (0.125) = 0.25$

$T_3(1) = 0.125$

$L_1(1) = (0.125)\ (2) = 0.25$

$L_2(1) = (0.250)\ (1) = 0.25$

$L_3(1) = (0.125)\ (4) = 0.50$

$P_3(1|1) = (4)\ (0.125)\ P_3(0|0) = 0.5$

$P_3(0|1) = 1 - P(1|1) = 0.5$.

Calculations for $v=2$:

$R_1(2) = (2)\ (1.25) = 2.5$

$R_2(2) = 1$

$R_3(2) = (1)(4)\ P_3(0|1) + (2)(2)\ P_3(1|1) = 4$

$$T_1(2) = \frac{(0.5)\ (2)}{(0.5)(2.5) + (1)(1) + (0.5)(4)} = 0.235$$

$T_2(2) = (2)\ (0.235) = 0.471$

$T_3(2) = (1)\ (0.235) = 0.235$

$$L_1(2) = (0.235)\ (2.5) = 0.588$$

$$L_2(2) = (0.471)\ (1) \quad = 0.471$$

$$L_3(2) = (0.235)\ (4) \quad = 0.940$$

$$P_3(1|2) = (4)\ (0.235)\ P_3(0|1) = 0.470$$

$$P_3(2|2) = (2)\ (0.235)\ P_3(1|1) = 0.235$$

$$P_3(0|2) = 1 - (0.470 + 0.235) = 0.295\ .$$

Calculations for v=3

$$R_1(3) = (2)\ (1.588) = 3.176$$

$$R_2(3) = 1$$

$$R_3(3) = (1)(4)\ P_3(0|2) + (2)(2)\ P_3(1|2) + (3)(2)\ P_3(2|2) = 4.470$$

$$T_1(3) = \frac{(0.5)\ (3)}{(0.5)(3.176) + (1)(1) + (0.5)(4.470)} = 0.311$$

$$T_2(3) = (2)\ (0.311) = 0.622$$

$$T_3(3) = (1)\ (0.311) = 0.311$$

$$L_1(3) = (0.311)\ (3.176) = 0.988$$

$$L_2(3) = (0.622)\ (1) \quad = 0.622$$

$$L_3(3) = (0.311)\ (4.470) = 1.390$$

$$P_3(1|3) = (4)(0.311)(0.295) = 0.367$$

$$P_3(2|3) = (2)(0.311)(0.470) = 0.292$$

$$P_3(3|3) = (2)(0.311)(0.235) = 0.146$$

$$P_3(0|3) = 1 - (0.367 + 0.292 + 0.146) = 0.195$$

$\rho_1(3) = (0.311)\ (2) = 0.622$

$\rho_2(3) = 0$

$\rho_3(3) = [(1)(0.377) + (2)(0.292) + (2)(0.146)]/2 = 0.622$.

As before, the throughput theorem can be used to determine the normalizing constant:

$G(0) = 1$,

$$G(1) = \frac{(1)(0.5)}{0.125} = 4 ,$$

$$G(2) = \frac{(4)(0.5)}{0.235} = 8.5$$

$$G(3) = \frac{(8.5)(0.5)}{0.311} = 13.667 .$$

The calculations for load dependent service centers not only require additional work, but also require additional storage to compute the marginal probabilities.

7.2.5 Load Independent - Multiple Chain - Closed Networks

The single chain MVA algorithm described in the previous two sections generalizes directly into a multiple chain algorithm. For Networks with load independent or type IS service centers, the recurrence equations are:

$$R_{ij}(V) = \begin{cases} s_{ij}\ [1 + L_i(V-1_j)] & \text{for } i \text{ FCFS, PS or LCFSPR} \\ s_{ij} & \text{for } i \text{ IS} \end{cases} \qquad (7.16)$$

$$T_{ij}(V) = e_{ij}\ v_j \ / \sum_{i=1}^{N} e_{ij}\ R_{ij}(V) \qquad (7.17)$$

$$L_{ij}(V) = T_{ij}(V)\ R_{ij}(V) \qquad (7.18)$$

$$L_i(V) = \sum_{j=1}^{J} L_{ij}(V)\ . \qquad (7.19)$$

The relationships expressed by the equations should be obvious from the earlier discussion with the possible exception of the equation for response time. That is, one might suspect that since customers in different chains may have different service rates if the discipline is PS or LCFSPR, the equations should somehow account for this. However, the equation is correct as stated, and the authors of the algorithm simply state that $s_{ij}\ L_i(V-1_j)$ is a congestion factor caused by the other customers. In the case of FCFS service centers, it is required that $s_{i1}=s_{i2}= \dots =s_{iJ}$ since MVA is only valid for networks that have product form solutions. A program-like definition of the algorithm is given in Table 7.3. Notice that the number of iterations has increased significantly due to multiple chains. As before, utilization can be calculated at any point in the procedure by the equations:

$$\rho_{ij}(V) = s_{ij}\ T_{ij}(V)\ , \qquad (7.20)$$

and

$$\rho_i(V) = \sum_{j=1}^{N} \rho_{ij}(V)\ . \qquad (7.21)$$

```
Begin {initialization
  for i:=1 to-N do $L_i(0)$ := 0
  for $v_1$:=0 to $V_1$ do
    for $v_2$:=0 to $V_2$ do
      ...
        for $v_J$:=0 to $V_J$ do
          begin {main body}
            $v = (v_1, v_2, \ldots, v_J)$
            for j:=1 to J do
              for i:=1 to N do {response time}
                $R_{ij}(v) = \begin{cases} s_{ij}\ [1 + L_i(v-1_j)] & \text{for i FCFS, PS, LCFSPR} \\ s_{ij} & \text{for i IS} \end{cases}$
              {throughput}
                $T_{1j}(v) = e_{1j}\ v_j \ / \sum_{i=1}^{N} e_{ij}\ R_{ij}(v)$
              for i:=2 to N do
                $T_{ij}(v) = (e_{ij}/e_{1j})\ T_{ij}(v)$
              for i:=1 to N do {chain queue length}
                $L_{ij}(v) = T_{ij}(v)\ R_{ij}(v)$
            end {for j}
            for i:=1 to N do {queue length}
              $L_i(v) = \sum_{j=1}^{J} L_{ij}(v)$
    end end end {for $v_1, v_2, \ldots, v_J$}
  {calculate utilization}
End.
```

Table 7.3 MVA Algorithm for Multiple Chain Load Independent Networks.

In order to illustrate the multiple chain procedure, consider the network in Figure 7.2 and the description of the service centers given in Table 7.4. Assigning $e_{3,1} = e_{3,2} = 1$ yields:

$$e_{1,1} = 0.5 \qquad e_{1,2} = 0.5$$
$$e_{2,1} = 0.5 \qquad e_{2,2} = 0.5$$
$$e_{3,1} = 1 \qquad e_{3,2} = 1 \ .$$

Thus,

$$T_{2,1}(V) = T_{1,1}(V) \qquad T_{2,2}(V) = T_{1,2}(V)$$
$$T_{3,1}(V) = (2)\ T_{1,1}(V) \qquad T_{3,2}(V) = (2)\ T_{1,2}(V),$$

for all V.

It follow from Table 7.4 that :

$$s_{1,1} = 2 \qquad s_{1,2} = 2$$
$$s_{2,1} = 4 \qquad s_{2,2} = 2$$
$$s_{3,1} = 1 \qquad s_{3,2} = 2 \ .$$

The following are the calculations up to $V=(2,2)$:

Calculations for $v = (0,1)$

$R_{1,2}(0,1) = 2$

$R_{2,2}(0,1) = 2$

$R_{3,2}(0,1) = 2$

$$T_{1,2}(0,1) = \frac{(0.5)\ (1)}{(0.5)(2)+(0.5)(2)+(1)(2)} = 0.125$$

$T_{2,2}(0,1) = (1)\ T_{1,2}(0,1) = 0.125$

$T_{3,2}(0,1) = (2)\ T_{1,2}(0,1) = 0.250$

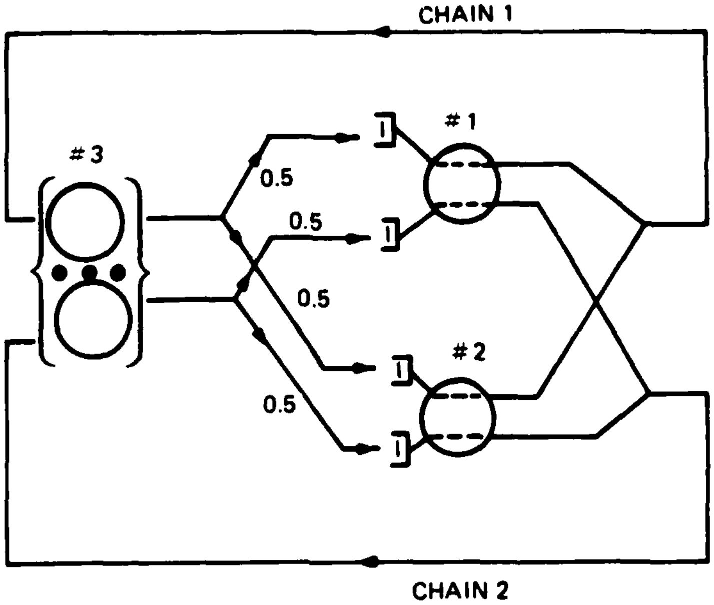


Figure 7.2 Load Independent, Multiple Chain Closed Network.

Service Center	Number of Servers	Service Discipline	Server Rate μ_{i1}	Server Rate μ_{i2}
1	1	FCFS	0.5	0.5
2	1	PS	0.25	0.5
3	∞	IS	1	0.5

Table 7.4 Description of Service Centers in Figure 7.2.

$$L_{1,2}(0,1) = (0.125)\ (2) = 0.25 = L_1(0,1)$$

$$L_{2,2}(0,1) = (0.125)\ (2) = 0.25 = L_3(0,1)$$

$$L_{3,2}(0,1) = (0.250)\ (2) = 0.50 = L_3(0,1)\ .$$

Calculations for v = (0,2) :

$$R_{1,2}(0,2) = (2)\ (1.25) = 2.5$$

$$R_{2,2}(0,2) = (2)\ (1.25) = 2.5$$

$$R_{3,2}(0,2) = 2$$

$$T_{1,2}(0,2) = \frac{(0.5)\ (2)}{(0.5)(2.5)+(0.5)(2.5)+(1)(2)} = 0.222$$

$$T_{2,2}(0,2) = (1)\ T_{1,2}(0,1) = 0.222$$

$$T_{3,2}(0,2) = (2)\ T_{1,2}(0,1) = 0.444$$

$$L_{1,2}(0,2) = (0.222)\ (2.5) = 0.556 = L_1(0,2)$$

$$L_{2,2}(0,2) = (0.222)\ (2.5) = 0.556 = L_2(0,2)$$

$$L_{3,2}(0,2) = (0.444)\ (2) \quad = 0.888 = L_3(0,2)\ .$$

Calculations for v = (1,0)

$$R_{1,1}(1,0) = (2)\ (1) = 2$$

$$R_{2,1}(1,0) = (4)\ (1) = 4$$

$$R_{3,1}(1,0) = 1$$

$$T_{1,1}(1,0) = \frac{(0.5)\ (2)}{(0.5)(2)+(0.5)(4)+(1)(1)} = 0.125$$

$$T_{2,1}(1,0) = 0.125$$

$$T_{3,1}(1,0) = 0.250$$

$$L_{1,1}(1,0) = (0.125)\ (2) = 0.25 = L_1(1,0)$$

$$L_{2,1}(1,0) = (0.125)\ (4) = 0.50 = L_2(1,0)$$

$$L_{3,1}(1,0) = (0.25)\ (1) = 0.25 = L_3(1,0)\ .$$

Calculations for $v = (1,1)$

$$R_{1,1}(1,1) = (2)\ (1.25) = 2.5$$

$$R_{2,1}(1,1) = (4)\ (1.25) = 5$$

$$R_{3,1}(1,1) = 1$$

$$T_{1,1}(1,1) = \frac{(0.5)\ (1)}{(0.5)(2.5)+(0.5)(5)+(1)(1)} = 0.105$$

$$T_{2,1}(1,1) = 0.105$$

$$T_{3,1}(1,1) = 0.210$$

$$L_{1,1}(1,1) = (0.105)\ (2.5) = (0.263)$$

$$L_{2,1}(1,1) = (0.105)\ (5) = (0.526)$$

$$L_{3,1}(1,1) = (0.210)\ (1) = (0.210)$$

$$R_{1,2}(1,1) = (2)\ (1.25) = 2.5$$

$$R_{2,2}(1,1) = (2)\ (1.5) = 3$$

$$R_{3,1}(1,1) = 2$$

$$T_{1,2}(1,1) = \frac{(0.5)\ (1)}{(0.5)(2.5)+(0.5)(3)+(1)(2)} = 0.105$$

$$T_{2,2}(1,1) = 0.105$$

$$T_{3,2}(1,1) = 0.211$$

$L_{1,2}(1,1) = (0.105)\ (2.5) = 0.263$

$L_{2,2}(1,1) = (0.105)\ (3) \quad = 0.315$

$L_{3,2}(1,1) = (0.211)\ (2) \quad = 0.422$

$L_{1}(1,1) = (0.263) + (0.263) = 0.526$

$L_{2}(1,1) = (0.526) + (0.315) = 0.841$

$L_{3}(1,1) = (0.210) + (0.422) = 0.632$.

Calculations for $v = (1,2)$:

$R_{1,1}(1,2) = (2)\ (1.556) = 3.112$

$R_{2,1}(1,2) = (4)\ (1.556) = 6.224$

$R_{3,1}(1,2) = 1$

$$T_{1,1}(1,2) = \frac{(0.5)\ (1)}{(0.5)(3.112)+(0.5)(6.224)+(1)(1)} = 0.088$$

$T_{2,1}(1,2) = 0.088$

$T_{3,1}(1,2) = 0.176$

$L_{1,1}(1,2) = (0.088)\ (3.112) = 0.274$

$L_{2,1}(1,2) = (0.088)\ (6.224) = 0.548$

$L_{3,1}(1,2) = (0.176)\ (1) \quad = 0.176$

$R_{1,2}(1,2) = (2)\ (1.526) = 3.052$

$R_{2,2}(1,2) = (2)\ (1.841) = 3.682$

$R_{3,2}(1,2) = 2$

$$T_{1,2}(1,2) = \frac{(0.5)\ (2)}{(0.5)(3.052)+(0.5)(3.682)+(1)(2)} = 0.186$$

$$T_{2,2}(1,2) = 0.186$$

$$T_{3,2}(1,2) = 0.373$$

$$L_{1,2}(1,2) = (0.186)\ (3.052) = 0.568$$

$$L_{2,2}(1,2) = (0.186)\ (3.682) = 0.685$$

$$L_{3,2}(1,2) = (0.373)\ (2) = 0.745$$

$$L_{1}(1,2) = (0.274) + (0.568) = 0.842$$

$$L_{2}(1,2) = (0.548) + (0.685) = 1.233$$

$$L_{3}(1,2) = (0.176) + (0.745) = 0.921\ .$$

Calculations for $v = (2,0)$:

$$R_{1,1}(2,0) = (2)\ (1.25) = 2.5$$

$$R_{2,1}(2,0) = (4)\ (1.5) = 6$$

$$R_{3,1}(2,0) = 1$$

$$T_{1,1}(2,0) = \frac{(0.5)\ (2)}{(0.5)(2.5)+(0.5)(6)+(1)(1)} = 0.190$$

$$T_{2,1}(2,0) = 0.190$$

$$T_{3,1}(2,0) = 0.381$$

$$L_{1,1}(2,0) = (0.190)\ (2.5) = 0.475 = L_{1}(2,0)$$

$$L_{2,1}(2,0) = (0.190)\ (6) = 1.143 = L_{2}(2,0)$$

$$L_{3,1}(2,0) = (0.381)\ (1) = 0.381 = L_{3}(2,0)\ .$$

Calculations for $v = (2,1)$:

$R_{1,1}(2,1) = (2)\ (1.526) = 3.052$

$R_{2,1}(2,1) = (4)\ (1.841) = 7.364$

$R_{2,1}(2,1) = 1$

$$T_{1,1}(2,1) = \frac{(0.5)\ (2)}{(0.5)(3.052)+(0.5)(7.364)+(1)(1)} = 0.161$$

$T_{2,1}(2,1) = 0.161$

$T_{3,1}(2,1) = 0.322$

$L_{1,1}(2,1) = (0.161)\ (3.052) = 0.492$

$L_{2,1}(2,1) = (0.161)\ (7.364) = 1.186$

$L_{3,1}(2,1) = (0.322)\ (1) \quad = 0.322$

$R_{1,2}(2,1) = (2)\ (1.475) = 2.950$

$R_{2,2}(2,1) = (2)\ (2.143) = 4.286$

$R_{3,2}(2,1) = 2$

$$T_{1,2}(2,1) = \frac{(0.5)\ (1)}{(0.5)(2.950)+(0.5)(4.286)+(1)(2)} = 0.089$$

$T_{2,2}(2,1) = 0.089$

$T_{3,2}(2,1) = 0.178$

$L_{1,2}(2,1) = (0.089)\ (2.950) = 0.263$

$L_{2,2}(2,1) = (0.089)\ (4.286) = 0.381$

$L_{3,2}(2,1) = (0.178)\ (1) \quad = 0.356$

$L_1(2,1) = 0.491 + 0.263 = 0.754$

$L_2(2,1) = 1.186 + 0.381 = 1.567$

$L_3(2,1) = 0.322 + 0.356 = 0.678$.

Calculations for $v = (2,2)$:

$R_{1,1}(2,2) = (2)\ (1.842) = 3.684$

$R_{2,1}(2,2) = (4)\ (2.233) = 8.932$

$R_{3,1}(2,2) = 1$

$$T_{1,1}(2,2) = \frac{(0.5)\ (2)}{(0.5)(3.684)+(0.5)(8.932)+(1)(1)} = 0.137$$

$T_{2,1}(2,2) = 0.137$

$T_{3,1}(2,2) = 0.274$

$L_{1,1}(2,2) = (0.137)\ (3.684) = 0.504$

$L_{2,1}(2,2) = (0.137)\ (8.932) = 1.222$

$L_{3,1}(2,2) = (0.274)\ (1) \quad = 0.274$

$R_{1,2}(2,2) = (2)\ (1.754) = 3.508$

$R_{2,2}(2,2) = (2)\ (2.567) = 5.134$

$R_{3,2)}(2,2) = 2$

$$T_{1,2}(2,2) = \frac{(0.5)\ (2)}{(0.5)(3.508)+(0.5)(5.134)+(1)(2)} = 0.158$$

$T_{2,2}(2,2) = 0.158$

$T_{3,2}(2,2) = 0.316$

$L_{1,2}(2,2) = (0.158)\ (3.508) = 0.555$

$L_{2,2}(2,2) = (0.158)\ (5.134) = 0.812$

$L_{3,2}(2,2) = (0.316)\ (2) \qquad = 0.633$

$L_1(2,2) = 0.504 + 0.555 = 1.059$

$L_2(2,2) = 1.222 + 0.812 = 2.034$

$L_3(2,2) = 0.274 + 0.633 = 0.907$.

Utilization Calculations :

$\rho_{1,1}(2,2) = (2)\ (0.137) = 0.274$

$\rho_{2,1}(2,2) = (4)\ (0.137) = 0.548$

$\rho_{3,1}(2,2) = 0$

$\rho_{1,2}(2,2) = (2)\ (0.158) = 0.316$

$\rho_{2,2}(2,2) = (2)\ (0.158) = 0.316$

$\rho_{3,2}(2,2) = 0$

$\rho_1(2,2) = 0.274 + 0.316 = 0.590$

$\rho_2(2,2) = 0.548 + 0.316 = 0.864$

$\rho_3(2,2) = 0$.

Normalizing Constant Calculations:

$$G(0,1) = \frac{(1)\ (0.5)}{0.125} = 4$$

$$G(0,2) = \frac{(4)\ (0.5)}{0.222} = 9.009$$

$$G(1,0) = \frac{(1)\ (0.5)}{0.125} = 4$$

$$G(1,1) = \frac{(4)\ (0.5)}{0.105} = 19.048$$

$$G(1,2) = \frac{(9.009)\ (0.5)}{0.088} = 51.188$$

$$G(2,0) = \frac{(4)\ (0.5)}{0.190} = 10.526$$

$$G(2,1) = \frac{(19.048)\ (0.5)}{0.161} = 59.1552$$

$$G(2,2) = \frac{(51.188)\ (0.5)}{0.137} = 186.818\ .$$

This problem demonstrates the primary reason why there are no examples of MVA in open literature, and very few examples of the other algorithms. Simply put, they are just too long! In addition, the calculations are iterative in nature and best done by a computer. The problem is that if one does not understand how to apply the algorithm, then they would not be able to write a computer program to do the calculations.

Appendix B is the listing of a computer program for multiple chain, load independent, closed networks. The code was written in Turbo Pascal, and is for an IBM PC or compatible computer. The program assumes that there is only one class of customers per chain. Therefore, if there are multiple classes per chain, the users must merge these into an equivalent class (See Chapter 6, Section 6.5.3 where customers are distinguished according to chains rather than class). In addition, the user must calculate the relative throughputs and supply them to the program. It is usually trivial to find the relative throughputs, and it was felt that this would be better than prompting the user for the routing probabilities (there are always more routing probabilities than throughputs).

7.2.6 Load Dependent - Multiple Chain - Closed Networks

As before, for service centers with limited load dependent service rates, the equation for the response time is in terms of marginal probabilities. More precisely,

$$R_{ij}(V) = \sum_{k_i=1}^{|V|} k_i \; s_{ij}(k_i) \; P_i(k_i-1|V-1_j) \; , \qquad (7.22)$$

where $|V| = V_1+V_2+ \cdots +V_J$, and

$$P_i(k_i|V) = \begin{cases} 1 & \text{for } k_i=0 \text{ and } |V|=0 \\ \sum_{j=1}^{J} s_{ij}(k_i) \; T_{ij}(V) \; P_i(k_i-1|V-1_j) & \text{for } k_i>0 \text{ and } |V|>0 \\ 1 - \sum_{k_i=1}^{|V|} P_i(k_i|V) & \text{for } k_i=0 \text{ and } |V|>0. \end{cases} \qquad (7.23)$$

Also as before, these same recursive equations apply to service centers that are load independent. In order to illustrate their use, a second server will be added to service center one of the previous multiple chain example and the problem reworked. The input parameters for the model are:

$$e_{1,1} = 0.5 \qquad e_{1,2} = 0.5$$

$$e_{2,1} = 0.5 \qquad e_{2,2} = 0.5$$

$$e_{3,1} = 1 \qquad e_{3,2} = 1$$

$$s_{1,2}(k_i) = \begin{cases} 2 \text{ for } k_i=1 \\ 1 \text{ for } k_i>1 \end{cases} \qquad s_{1,2}(k_i) = \begin{cases} 2 \text{ for } k_i=1 \\ 1 \text{ for } k_i>1 \end{cases}$$

$$s_{2,2} = 4 \qquad s_{2,2} = 2$$

$$s_{3,2} = 1 \qquad s_{3,2} = 2$$

$$T_{2,1}(V) = (e_{2,1}/e_{1,1})\ T_{1,1}(V) = (1)\ T_{1,1}(V)$$

$$T_{3,1}(V) = (e_{3,1}/e_{1,1})\ T_{1,1}(V) = (2)\ T_{1,1}(V)$$

$$T_{2,2}(V) = (e_{2,2}/e_{1,2})\ T_{1,2}(V) = (1)\ T_{1,2}(V)$$

$$T_{3,2}(V) = (e_{3,2}/e_{1,2})\ T_{1,2}(V) = (2)\ T_{1,2}(V)\ .$$

Calculations for $v = (0,1)$:

$$R_{1,2}(0,1) = (1)\ s_{1,2}(1)\ P_1(0|0,0)$$

$$= (1)\ (2)\ (1) = 2$$

$$R_{2,2}(0,1) = (2)\ (1) = 2$$

$$R_{3,2}(0,1) = 2$$

$$T_{1,2}(0,1) = \frac{(0.5)\ (1)}{(0.5)(2) + (0.5)(2) + (1)(2)} = 0.125$$

$$T_{2,2}(0,1) = (1)\ (0.125) = 0.125$$

$$T_{3,3}(0,1) = (2)\ (0.125) = 0.250$$

$$L_{1,2}(0,1) = (0.125)\ (2) = 0.25 = L_1(0,1)$$

$$L_{2,1}(0,1) = (0.125)\ (2) = 0.25 = L_2(0,1)$$

$$L_{3,1}(0,1) = (0.250)\ (2) = 0.50 = L_3(0,1)$$

$$\begin{aligned} P_1(1|0,1) &= s_{1,2}(1)\ T_{1,2}(0,1)\ P_1(0|0,0) \\ &= (2)\ (0.125)\ (1) = 0.25 \end{aligned}$$

$$P_1(0|0,1) = 1 - 0.25 = 0.75.$$

Calculations for $v = (0,2)$:

$$\begin{aligned} R_{1,2}(0,2) &= (1)\ s_{1,2}(1)\ P_1(0|0,1) + (2)\ s_{1,2}(2)\ P_1(1|0,1) \\ &= (1)\ (2)\ (0.75) + (2)\ (1)\ (0.25) = 2 \end{aligned}$$

$$R_{2,2}(0,2) = (2)\ (1.25) = 2.5$$

$$R_{3,2}(0,2) = 2$$

$$T_{1,2}(0,2) = \frac{(0.5)\ (2)}{(0.5)(2) + (0.5)(2.5) + (1)(2)} = 0.235$$

$$T_{2,2}(0,2) = (1)\ (0.235) = 0.235$$

$$T_{3,2}(0,2) = (2)\ (0.235) = 0.471$$

$$L_{1,2}(0,2) = (0.235)\ (2) \quad = 0.471 = L_1(0,2)$$

$$L_{2,2}(0,2) = (0.235)\ (2.5) = 0.588 = L_2(0,2)$$

$$L_{3,2}(0,2) = (0.471)\ (2) \quad = 0.941 = L_3(0,2)$$

$$\begin{aligned} P_1(1|0,2) &= s_{1,2}(1)\ T_{1,2}(0,2)\ P_1(0|0,1) \\ &= (1)\ (0.235)\ (0.25) = 0.353 \end{aligned}$$

$$\begin{aligned} P_1(2|0,2) &= s_{1,2}(2)\ T_{1,2}(0,2)\ P_1(1|0,1) \\ &= (1)\ (0.235)\ (0.25) = 0.059 \end{aligned}$$

$$P_1(0|0,2) = 1 - (0.353 + 0.059) = 0.588\ .$$

Calculations for $v = (1,0)$:

$$R_{1,1}(1,0) = (1)\ s_{1,1}(1)\ P_1(0|0,0)$$

$$= (1)\ (2)\ (1) = 2$$

$$R_{2,1}(1,0) = (4)\ (1) = 4$$

$$R_{3,1}(1,0) = 1$$

$$T_{1,1}(1,0) = \frac{(0.5)\ (1)}{(0.5)(2) + (0.5)(4) + (1)(1)} = 0.125$$

$$T_{2,1}(1,0) = (1)\ (0.125) = 0.125$$

$$T_{3,1}(1,0) = (2)\ (0.125) = 0.250$$

$$L_{1,1}(1,0) = (0.125)\ (2) = 0.25 = L_1(1,0)$$

$$L_{2,1}(1,0) = (0.125)\ (4) = 0.50 = L_2(1,0)$$

$$L_{3,1}(1,0) = (0.250)\ (1) = 0.25 = L_3(1,0)$$

$$P_1(1|1,0) = s_{1,1}\ T_{1,1}(1,0)\ P_1(0|0,0)$$

$$= (2)\ (0.125)\ (1) = 0.25$$

$$P_1(0|1,0) = 1 - 0.25 = 0.75\ .$$

Calculations for $v = (1,1)$:

$$R_{1,1}(1,1) = (1)\ s_{1,1}(1)\ P_1(0|0,1) + (2)\ s_{1,1}(2)\ P_1(1|0,1)$$

$$= (1)\ (2)\ (0.75) + (2)\ (1)\ (0.25) = 2$$

$$R_{2,1}(1,1) = (4)\ (1.25) = 5$$

$$R_{3,1}(1,1) = 1$$

$$T_{1,1}(1,1) = \frac{(0.5)\ (1)}{(0.5)(2) + (0.5)(5) + (1)(1)} = 0.111$$

$$T_{2,1}(1,1) = (1)\ (0.111) = 0.111$$

$$T_{3,1}(1,1) = (2)\ (0.111) = 0.222$$

$$L_{1,1}(1,1) = (0.111)\ (2) = 0.222$$

$$L_{2,1}(1,1) = (0.111)\ (5) = 0.555$$

$$L_{3,1}(1,1) = (0.222)\ (1) = 0.222$$

$$R_{1,2}(1,1) = (1)\ s_{1,2}(1)\ P_1(0|1,0) + (2)\ s_{1,2}(2)\ P_1(1|1,0)$$

$$= (1)(2)(0.75) + (2)(1)(0.25) = 2$$

$$R_{2,2}(1,1) = (2)\ (1.5) = 3$$

$$R_{3,2}(1,1) = 2$$

$$T_{1,2}(1,1) = \frac{(0.5)(1)}{(0.5)(2) + (0.5)(3) + (1)(2)} = 0.111$$

$$T_{2,2}(1,1) = (1)\ (0.111) = 0.111$$

$$T_{3,2}(1,1) = (2)\ (0.111) = 0.222$$

$$L_{1,2}(1,1) = (0.111)\ (2) = 0.222$$

$$L_{2,2}(1,1) = (0.111)\ (3) = 0.333$$

$$L_{3,2}(1,1) = (0.222)\ (2) = 0.444$$

$$L_1(1,1) = 0.222 + 0.222 = 0.444$$

$$L_2(1,1) = 0.555 + 0.333 = 0.888$$

$$L_3(1,1) = 0.222 + 0.444 = 0.666$$

$P_1(1|1,1) = s_{1,1}(1)\ T_{1,1}(1,1)\ P_1(0|0,1) + s_{1,2}(1)\ T_{1,2}(1,1)\ P_1(0|1,0)$

$\quad = (2)(0.111)(0.75) + (2)(0.111)(0.75) = 0.333$

$P_1(2|1,1) = s_{1,1}(2)\ T_{1,1}(1,1)\ P_1(1|0,1) + s_{1,2}(2)\ T_{1,2}(1,1)\ P_1(1|1,0)$

$\quad = (1)(0.111)(0.25) + (1)(0.111)(0.25) = 0.056$

$P_1(0|1,1) = 1 - (0.333 + 0.056) = 0.611$.

Calculations for $v = (1,2)$:

$R_{1,1}(1,2) = (1)\ s_{1,1}(1)\ P_1(0|0,2) + (2)\ s_{1,1}(2)\ P_1(1|0,2)$

$\quad + (3)\ s_{1,1}(3)\ P_1(2|0,2)$

$\quad = (1)(2)(0.588) + (2)(1)(0.353) + (3)(1)(0.059) = 2.059$

$R_{2,1}(1,2) = (4)\ (1.588) = 6.352$

$R_{3,1}(1,2) = 1$

$$T_{1,1}(1,2) = \frac{(0.5)(1)}{(0.5)(2.059) + (0.5)(6.353) + (1)(1)} = 0.096$$

$T_{2,1}(1,2) = (1)\ (0.096) = 0.096$

$T_{3,1}(1,2) = (2)\ (0.096) = 0.192$

$L_{1,1}(1,2) = (0.096)\ (2.059) = 0.198$

$L_{2,1}(1,2) = (0.096)\ (6.352) = 0.610$

$L_{3,1}(1,2) = (0.192)\ (1) \quad = 0.192$

$R_{1,2}(1,2) = (1)\ s_{1,2}(1)\ P_1(0|1,1) + (2;\ s_{1,2}(2)\ P_1(1|1,1)$

$\quad + (3)\ s_{1,2}(3)\ P_1(2|1,1)$

$\quad = (1)(2)(0.611) + (2)(1)(0.333) + (3)(1)(0.056) = 2.056$

$R_{2,2}(1,2) = (2)(1.888) = 3.776$

$R_{3,2}(1,2) = 2$

$$T_{1,2}(1,2) = \frac{(0.5)(2)}{(0.5)(2.056) + (0.5)(3.776) + (1)(2)} = 0.203$$

$$T_{2,2}(1,2) = (1)\ (2.03) = 0.203$$

$$T_{3,2}(1,2) = (2)\ (2.03) = 0.407$$

$$L_{1,2}(1,2) = (0.203)\ (2.056) = 0.417$$

$$L_{2,2}(1,2) = (0.203)\ (3.776) = 0.767$$

$$L_{3,2}(1,2) = (0.407)\ (2) \quad = 0.814$$

$$L_1(1,2) = 0.198 + 0.417 = 0.615$$

$$L_2(1,2) = 0.610 + 0.767 = 1.377$$

$$L_3(1,2) = 0.192 + 0.814 = 1.006$$

$$\begin{aligned} P_1(1|1,2) &= s_{1,1}(1)\ T_{1,1}(1,2)\ P_1(0|0,2) + s_{1,2}(1)\ T_{1,2}(1,2)\ P_1(0|1,1) \\ &= (2)(0.096)(0.588) + (2)(.203)(0.611) = 0.361 \end{aligned}$$

$$\begin{aligned} P_1(2|1,2) &= s_{1,1}(2)\ T_{1,1}(1,2)\ P_1(1|0,2) + s_{1,2}(2)\ T_{1,2}(1,2)\ P_1(1|1,1) \\ &= (1)(0.096)(0.535) + (1)(.203)(0.333) = 0.101 \end{aligned}$$

$$\begin{aligned} P_1(3|1,2) &= s_{1,1}(3)\ T_{1,1}(1,2)\ P_1(2|0,2) + s_{1,2}(3)\ T_{1,2}(1,2)\ P_1(2|1,1) \\ &= (1)(0.096)(0.059) + (1)(.203)(0.056) = 0.017 \end{aligned}$$

$$P_1(0|1,2) = 1 - (0.361 + 0.101 + 0.017) = 0.521\ .$$

Calculations for $v=(2,0)$

$$\begin{aligned} R_{1,1}(2,0) &= (1)\ s_{1,1}(1)\ P_1(0|1,0) + (2)\ s_{1,1}(2)\ P_1(1|1,0) \\ &= (1)(2)(0.75) + (2)(1)(0.25) = 2 \end{aligned}$$

$$R_{2,1}(2,0) = (4)(1.5) = 6$$

$$R_{3,1}(2,0) = 1$$

$$T_{1,1}(2,0) = \frac{(0.5)(2)}{(0.5)(2)+(0.5)(6)+(1)(1)} = 0.20$$

$$T_{2,1}(2,0) = (1)(0.20) = 0.20$$

$$T_{3,1}(2,0) = (2)(0.20) = 0.40$$

$$L_{1,1}(2,0) = (0.20)(2) = 0.40 = L_1(2,0)$$

$$L_{2,1}(2,0) = (0.20)(6) = 1.20 = L_2(2,0)$$

$$L_{3,1}(2,0) = (0.40)(1) = 0.40 = L_3(2,0)$$

$$\begin{aligned} P_1(1|2,0) &= s_{1,1}(1)\ T_{1,1}(2,0)\ P_1(0|1,0) \\ &= (2)(0.20)(0.75) = 0.300 \end{aligned}$$

$$\begin{aligned} P_1(2|2,0) &= s_{1,1}(1)\ T_{1,1}(2,0)\ P_1(1|1,0) \\ &= (1)(0.20)(0.25) = 0.050 \end{aligned}$$

$$P_1(0|2,0) = 1 -(0.300 + 0.050) = 0.650\ .$$

Calculations for $v=(2,1)$

$$\begin{aligned} R_{1,1}(2,1) &= (1)\ s_{1,1}(1)\ P_1(0|1,1) + (2)\ s_{1,1}(2)\ P_1(1|1,1) \\ &\quad + (3)\ s_{1,1}(3)\ P_1(2|1,1) \\ &= (1)(2)(0.611) + (2)(1)(0.333) + (3)(1)(0.056) = 2.056 \end{aligned}$$

$$R_{2,1}(2,1) = (4)(1.888) = 7.552$$

$$R_{3,1}(2,1) = 1$$

$$T_{1,1}(2,1) = \frac{(0.5)(2)}{(0.5)(2.056)+(0.5)(7.552)(1)(1)} = 0.172$$

$$T_{2,1}(2,1) = (1)\ (0.172) = 0.172$$

$$T_{3,1}(2,1) = (2)\ (0.172) = 0.345$$

$L_{1,1}(2,1) = (0.172)\ (2.056) = 0.354$

$L_{2,1}(2,1) = (0.172)\ (7.552) = 1.299$

$L_{3,1}(2,1) = (0.345)\ (1) \quad = 0.345$

$$\begin{aligned} R_{1,2}(2,1) &= (1)\ s_{1,2}(1)\ P_1(0|2,0) + (2)\ s_{1,2}(2)\ P_1(1|2,0) \\ &+ (3)\ s_{1,2}(3)\ P_1(2|2,0) \\ &= (1)(2)(0.650) + (2)(1)(0.300) + (3)(1)(0.050) = 2.050 \end{aligned}$$

$R_{2,2}(2,1) = (2)\ (2.20) = 4.40$

$R_{3,2}(2,1) = 2$

$$T_{1,1}(2,1) = \frac{(0.5)(1)}{(0.5)(2.050)+(0.5)(4.40)+(1)(2)} = 0.096$$

$T_{2,2}(2,1) = (1)\ (0.096) = 0.096$

$T_{3,2}(2,1) = (2)\ (0.096) = 0.191$

$L_{1,2}(2,1) = (0.096)\ (2.050) = 0.196$

$L_{2,2}(2,1) = (0.096)\ (4.40) \quad = 0.421$

$L_{3,2}(2,1) = (0.191)\ (2) \quad = 0.383$

$L_1(2,1) = 0.354 + 0.196 = 0.550$

$L_2(2,1) = 1.299 + 0.421 = 1.720$

$L_3(2,1) = 0.345 + 0.383 = 0.728$

$$\begin{aligned} P_1(1|1,2) &= s_{1,1}(1)\ T_{1,1}(2,1)\ P_1(0|1,1) + s_{1,2}(2)\ T_{1,2}(2,1)\ P_1(0|2,0) \\ &= (2)(0.172)(0.333) + (1)(0.096)(0.650) = 0.335 \end{aligned}$$

$$\begin{aligned} P_1(2|2,1) &= s_{1,1}(2)\ T_{1,1}(2,1)\ P_1(1|1,1) + s_{1,2}(2)\ T_{1,2}(2,1)\ P_1(1|2,0) \\ &= (1)(0.172)(0.333) + (1)(0.096)(0.300) = 0.086 \end{aligned}$$

$P_1(3|2,1) = s_{1,1}(3)\ T_{1,1}(2,1)\ P_1(2|1,1) + s_{1,2}(3)\ T_{1,2}(2,1)\ P_1(2|2,0)$

$\qquad = (1)(0.172)(0.056) + (1)(0.096)(0.050) = 0.014$

$P_1(0|2,1) = 1 - (0.335 + 0.0866 + 0.014) = 0.565$.

Calculations for $v=(2,2)$

$R_{1,1}(2,2) = (1)\ s_{1,1}(1)\ P_1(0|1,2) + (2)\ s_{1,1}(2)\ P_1(1|1,2)$

$\qquad + (3)\ s_{1,1}(3)\ P_1(2|1,2) + (4)\ s_{1,1}(4)\ P_1(3|1,2)$

$\qquad = (1)(2)(0.521)+(2)(1)(0.361)+(3)(1)(0.101)+(4)(1)(0.17)$

$\qquad = 2.135$

$R_{2,1}(2,2) = (4)\ (2.377) = 9.508$

$R_{3,1}(2,2) = 1$

$$T_{1,1}(2,2) = \frac{(0.5)(2)}{(0.5)(2.135)+(0.5)(9.508)+(1)(1)} = 0.147$$

$T_{2,1}(2,2) = (1)\ (0.147) = 0.147$

$T_{3,1}(2,2) = (2)\ (0.147) = 0.293$

$L_{1,1}(2,2) = (0.147)\ (2.135) = 0.313$

$L_{2,1}(2,2) = (0.147)\ (9.508) = 1.394$

$L_{3,1}(2,2) = (0.293)\ (1) \quad = 0.293$

$R_{1,2}(2,2) = (1)\ s_{1,2}(1)\ P_1(0|2,1) + (2)\ s_{1,2}(2)\ P_1(1|2,1)$

$\qquad = (3)\ s_{1,2}(3)\ P_1(2|2,1) + (4)\ s_{1,2}(4)\ P_1(3|2,1)$

$R_{2,2}(2,2) = (2)(2.720) = 5.440$

$R_{3,2}(2,2) = 2$

$$T_{1,2}(2,2) = \frac{(0.5)(2)}{[(0.5)(2.114)+(0.5)(5.440)+(1)(2)} = 0.173$$

$$T_{2,2}(2,2) = (1)\ (0.173) = 0.173$$

$$T_{3,2}(2,2) = (2)\ (0.173) = 0.346$$

$$L_{1,2}(2,2) = (0.173)\ (2.114) = 0.366$$

$$L_{2,2}(2,2) = (0.173)\ (5.440) = 0.942$$

$$L_{3,2}(2,2) = (0.313)\ (2) \quad = 0.692$$

$$L_1(2,2) = 0.313 + 0.366 = 0.679$$

$$L_2(2,2) = 1.394 + 0.942 = 2.336$$

$$L_3(2,2) = 0.293 + 0.692 = 0.985$$

$$\begin{aligned}P_1(1|2,2) &= s_{1,1}(1)\ T_{1,1}(2,2)\ P_1(0|1,2) + s_{1,2}(1)\ T_{1,2}(2,2)\ P_1(0|2,1)\\ &= (2)(0.147)(0.521) + (2)(0.173)(0.565) = 0.349\end{aligned}$$

$$\begin{aligned}P_1(2|2,2) &= s_{1,1}(2)\ T_{1,1}(2,2)\ P_1(1|1,2) + s_{1,2}(2)\ T_{1,2}(2,2)\ P_1(1|2,1)\\ &= (1)(0.147)(0.361) + (1)(0.173)(0.335) = 0.111\end{aligned}$$

$$\begin{aligned}P_1(3|2,2) &= s_{1,1}(3)\ T_{1,1}(2,2)\ P_1(2|1,2) + s_{1,2}(3)\ T_{1,2}(2,2)\ P_1(2|2,1)\\ &= (1)(0.147)(0.101) + (1)(0.173)(0.086) = 0.030\end{aligned}$$

$$\begin{aligned}P_1(4|2,2) &= s_{1,1}(4)\ T_{1,1}(2,2)\ P_1(3|1,2) + s_{1,2}(4)\ T_{1,2}(2,2)\ P_1(3|2,1)\\ &= (1)(0.147)(0.017) + (1)(0.173)(0.014) = 0.005\end{aligned}$$

$$P_1(4|2,2) = 1 - (0.349 + 0.111 + 0.030 + 0.005) = 0.505\ .$$

Utilization :

$$\rho_1(2,2) = [(1)(0.349)+(2)(0.111)+(3)(0.030)+(4)(0.005)]/2 = 0.341$$

$$\rho_2(2,2) = (4)(0.147) + (2)(0.173) = 0.934$$

$$\rho_3(2,2) = 0\ .$$

Normalizing Constants:

$$G(0,1) = \frac{\cancel{(1)}(0.5)}{0.125} = 4$$

$$G(0,2) = \frac{(4)(0.5)}{0.235} = 8.511$$

$$G(1,0) = \frac{(1)(0.5)}{0.125} = 4$$

$$G(1,1) = \frac{(4)(0.5)}{0.111} = 18.018$$

$$G(1,2) = \frac{(8.511)(0.5)}{0.096} = 44.328$$

$$G(2,0) = \frac{(4)(0.5)}{0.2} = 10$$

$$G(2,1) = \frac{(18.018)(0.5)}{0.172} = 52.378$$

$$G(2,2) = \frac{(44.328)(0.5)}{0.147} = 150.776\ .$$

As previously stated, the examples in this chapter are believed to be the only examples of MVA in open literature. In view of the length of this example, it should not be surprising that it is believed to be the only multiple chain, load dependent example-period.

7.3 Mixed Networks

A mixed network is one with both open and closed chains. Customers in the closed chains and their service requirement, do not affect the stability of the network [REIS75]. That is, a mixed network is stable, if and only if, $\rho_{i,open} < 1$, for all i other than IS service centers ($\rho_{i,open}$ is the utilization of service center i due to the open chain customers). The open and closed chains have a surprisingly simple and limited impact on each other. In fact, if the mean arrival rates of the open chains are constant and all service rates are load independent, then the response time is given by

$$R_{ij}(V) = \frac{s_{ij}\ [1 + L_i(V-1_j]}{1 - \rho_{i,open}} \qquad (7.24)$$

where $L_i(V-1_j)$ is the mean number of closed chain customers when the population vector is $V-1_j$ [ZAHO81]. Similarly, the response time for open chains is

$$R_{ij}(V) = \frac{s_{ij}\ [1 + L_i(V)]}{1 - \rho_{i,open}} \cdot \qquad (7.25)$$

The reader is referred to the earlier reference for a proof of these two equations. Note that it is only necessary to compute the open chain utilizations in order to determine closed chain metrics. Once the closed chain metrics have been determined, those of the open chain can be calculated. The procedure is best illustrated by an example. Figure 7.3 depicts a mixed network with two service centers. The specifications of the network are given in Table 7.5. The following calculations assume an arrival rate of 0.3 and two customers in the closed chain.

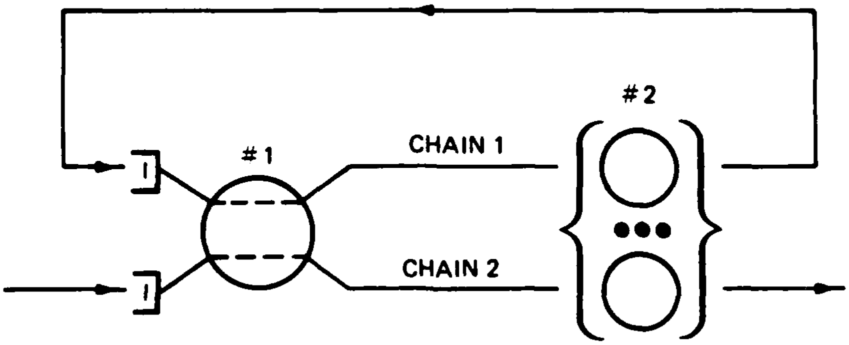


Figure 7.3 Example of a Mixed Network.

Service Center	Number of Servers	Service Discipline	Server Rate μ_{i1}	Server Rate μ_{i2}
1	1	FCFS	1/2	1/2
2	∞	IS	1/3	1/10

Table 7.5 Description of Service Centers in Figure 7.3.

Calculations:

$$\lambda_{1,2} = 0.3$$

$$s_{1,1} = 2 \qquad e_{1,1} = 1$$

$$s_{2,1} = 3 \qquad e_{2,1} = 1$$

$$s_{1,2} = 2 \qquad e_{1,2} = 1$$

$$s_{2,2} = 10 \qquad e_{2,2} = 1 .$$

Obviously the throughput of the open chain is equal to the arrival rate. Hence,

$$T_{1,2} = T_{2,2} = 0.3 .$$

The utilization at service center 1 due to the open chain is:

$$\rho_{1,2} = (0.3)(2) = 0.6 .$$

The performance metrics of the closed chain can now be determined.

Calculations for v = 1 :

$$R_{1,1}(1) = \frac{(2)(1)}{1 - 0.6} = 5$$

$$R_{2,1}(1) = 3$$

$$T_{1,1}(1) = \frac{(1)(1)}{(1)(5) + (1)(3)} = 0.125$$

$$T_{2,1}(1) = 0.125$$

$$L_{1,1}(1) = (0.125)\ (5) = 0.625$$

$$L_{2,1}(1) = (0.125)\ (3) = 0.375 .$$

Calculations for $v = 2$:

$$R_{1,1}(2) = \frac{(2)\ (1.625)}{1 - 0.6} = 8.125$$

$$R_{2,1}(2) = 3$$

$$T_{1,1}(2) = \frac{(1)\ (2)}{(1)(8.125) + (1)(3)} = 0.180$$

$$T_{2,1}(2) = 0.180$$

$$L_{1,1}(2) = (0.180)\ (8.125) = 1.461$$

$$L_{2,1}(2) = (0.180)\ (3) = 0.539$$

$$\rho_{1,1} = (0.180)(2) = 0.36\ .$$

Now that the metrics of the closed chain have been determined, those of the open chain can be calculated as follows:

$$R_{1,2}(2) = \frac{(2)\ (2.461)}{1 - 0.6} = 12.305$$

$$R_{2,2}(2) = s_{2,2} = 10$$

$$L_{1,2}(2) = (0.3)\ (12.305) = 3.692$$

$$L_{2,2}(2) = (0.3)\ (10) = 3\ .$$

The normalizing constant of a mixed network can be expressed as the product of two smaller normalizing constants, one for the open chains and the other for the closed chains. It follows from the previous chapter that the open chain normalizing constant is:

$$G_{(open)} = \prod_{i=1}^{N} G_i \ ,$$

$$\text{where } G_i = \begin{cases} \left[1 - \sum_c \lambda(e_{ic}/\mu_{ic})\right]^{-1} & \text{for i FCFS, PS, or LCFS} \\ \exp\left[\sum_c \lambda(e_{ic}/\mu_{ic})\right] & \text{for i IS.} \end{cases} \tag{7.26}$$

The normalizing constant of the closed chains can be determined from the throughput theorem as before, or both normalizing constants can be merged into one by observing that G(v) is directly proportional to G(0) for all v. Thus, the two normalizing constants can be combined into one by simply defining G(0) as $G_{(open)}$. The procedure will be illustrated by finding the normalizing constant in the previous example.

$$G_{(open)} = \frac{1}{1-0.6} e^3 = 50.214$$

$$G(0) = G_{(open)}$$

$$G(1) = \frac{G(0)}{T_{1,1}(1)} e_{1,1} = \frac{(50.214)(1)}{0.125} = 401.711$$

$$G(2) = \frac{G(1)}{T_{1,1}(2)} e_{1,1} = \frac{(401.711)(1)}{0.125} = 2.231.726 \ .$$

As before, it is simple to extend the equations for multiple chain, load independent networks. However, the limited load dependent case is considerably more complicated even for single chain networks, and will not be discussed here [REIS83]. In fact, the equations will not even be listed for reference purposes because this would require redefining the notation that has previously been used and would only complicate matters.

7.4 Closed Queueing Networks without Product Form Solutions

A major advantage in the analysis of closed queueing networks over open and mixed networks is that if the network can be represented by a 'pure' Markov process then, theoretically, a solution can be obtained. That is, since the number of customers is finite, the number of network states is finite, and therefore the process can be completely described by a finite set of linear equations which equate the rate of flow into a network state to the rate of flow out of the same state. Thus, the set of linear equations can be solved to obtain steady-state probabilities, and the other performance metrics can be obtained from these.

The procedure will be illustrated by an example. Consider the two service center network in Figure 7.6. It is assumed that service center 1 contains an infinite number of servers and that the service center 2 contains a single server and the service discipline is nonpreemptive priority. It is also assumed that all service times are exponentially distributed and that the number of customers in each of the three chains is one. Since there is only one customer per chain, let the numbers 1,2, and 3 represent these customers and let the network state

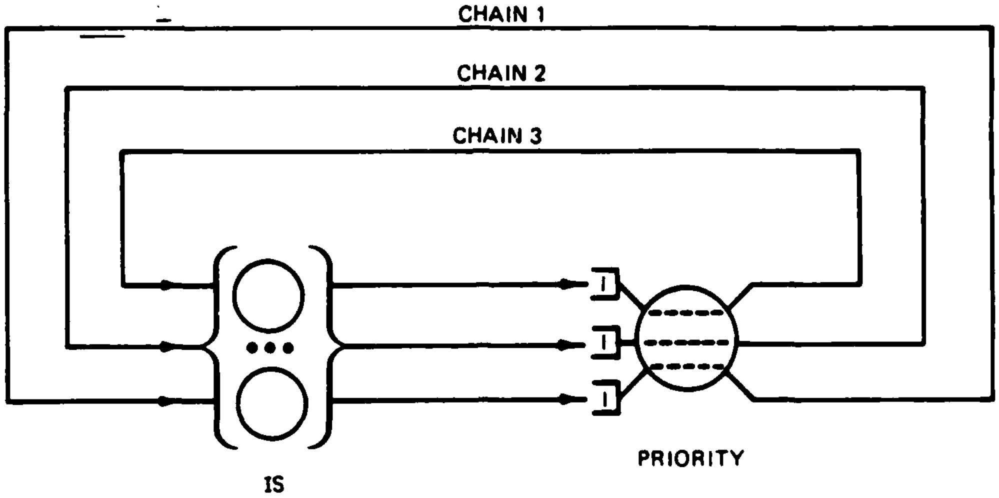


Figure 7.6 Closed Network Without a Product Form Solution.

be defined by P[(x)(y)] where (x) and (y) specify the customers at service centers 1 and 2, respectfully. Now since service center 1 contains an infinite number of servers, the order of customers at this service center is not important, however since service center 2 has only a single server and its service discipline is nonpreemptive priority, the specification order is important and must be included in its state specification. For example, P[(2),(3,1)] represents the network state where service center 1 contains the chain 2 customers and service center 2 contains the chain 3 and chain 1 customers. Note that the chain 3 customer is currently being served at service center 2. Also note that P[(0),(2,3,1)] is not a legitimate network state because when the chain 2 customer finishes service, the chain 1 customer will be served before the chain 3 customer. The steady-state equations describing the network are:

rate out = rate in

$$\mu_{21}P[(0)(1,2,3)] = \mu_{12}P[(2)(1,3)] + \mu_{13}P[(3)(1,2)]$$

$$\mu_{22}P[(0)(2,1,3)] = \mu_{11}P[(1)(2,3)] + \mu_{13}P[(3)(2,1)]$$

$$\mu_{23}P[(0)(3,1,2)] = \mu_{11}P[(1)(3,2)] + \mu_{12}P[(2)(3,1)]$$

$$(\mu_{21}+\mu_{13})P[(3)(1,2)] = \mu_{12}P[(2,3)(1)] + \mu_{23}P[(0)(3,1,2)]$$

$$(\mu_{21}+\mu_{12})P[(2)(1,3)] = \mu_{13}P[(2,3)(1)] + \mu_{22}P[(0)(2,1,3)]$$

$$(\mu_{22}+\mu_{13})P[(3)(2,1)] = \mu_{11}P[(1,3)(2)]$$

$$(\mu_{22}+\mu_{11})P[(1)(2,3)] = \mu_{13}P[(1,3)(2)] + \mu_{21}P[(0)(1,2,3)]$$

$$(\mu_{23}+\mu_{12})P[(2)(3,1)] = \mu_{11}P[(1,2)(3)]$$

$$(\mu_{23}+\mu_{11})P[(1)(3,2)] = \mu_{12}P[(1,2)(3)]$$

$$(\mu_{21}+\mu_{12}+\mu_{13})P[(2,3)(1)] = \mu_{11}P[(1,2,3)(0)] + \mu_{22}P[(3)(2,1)] + \mu_{23}P[(2)(3,1)]$$

$$(\mu_{22}+\mu_{11}+\mu_{13})P[(1,3)(2)] = \mu_{12}P[(1,2,3)(0)] + \mu_{21}P[(3)(1,2)] + \mu_{23}P[(1)(3,2)]$$

$$(\mu_{23}+\mu_{11}+\mu_{12})P[(1,2)(3)] = \mu_{13}P[(1,2,3)(0)] + \mu_{21}P[(2)(1,3)] + \mu_{22}P[(1)(2,3)]$$

$$(\mu_{11}+\mu_{12}+\mu_{13})P[(1,2,3)(0)] = \mu_{21}P[(2,3)(1)] + \mu_{22}P[(1,3)(2)] + \mu_{23}P[(1,2)(3)] \; .$$

Several key issues about these equations need to be emphasized. First and most significant is that local balance does not apply, and therefore, the procedure used to obtain the solution of networks with product form does not apply (nor can it be extended to do so!!!).

Secondly, there is no simple equation, as in the case of networks with product form solutions, to find the number of feasible networks states. That is, there is no easy way to determine the number of equations necessary to fully describe the process. Finally, the equations are less symmetrical than those of networks with product form solutions, and thus are considerably more complicated to write. Fortunately, however, there is a way to check them. If the equations are put into matrix form, then each column must sum to zero. The reason for this is that the set of equations are dependent. To obtain an independent set, one of the equations (any one) is replaced with conservation of probability equation (summation over all probabilities equal one).

Returning to the problem being considered, once the steady-state probabilities have been determined all other performance metrics can be found. Performance metrics by class can be obtained from the steady-state probabilities via the following equations:

$$L_{1,1} = P[(1)(2,3)] + P[(1)(3,2)] + P[(1,3)(2)] + P[(1,2)(3)] + P[(1,2,3)(0)]$$

$$L_{1,2} = P[(2)(1,3)] + P[(2)(3,1)] + P[(2,3)(1)] + P[(1,2)(3)] + P[(1,2,3)(0)]$$

$$L_{1,3} = P[(3)(1,2)] + P[(3)(2,1)] + P[(2,3)(1)] + P[(1,3)(2)] + P[(1,2,3)(0)]$$

$$L_{2,1} = P[(0)(1,2,3)] + P[(0)(2,1,3)] + P[(0)(3,1,2)] + P[(3)(1,2)] + P[(2)(1,3)] + P[(3)(2,1)] + P[(2)(3,1)] + P[(2,3)(1)]$$

$$L_{2,2} = P[(0)(1,2,3)] + P[(0)(2,1,3)] + P[(0)(3,1,2)] + P[(3)(1,2)]$$

$$+ P[(3)(2,1)] + P[(1)(2,3)] + P[(1)(3,2)] + P[(1,3)(2)]$$

$$L_{2,3} = P[(0)(1,2,3)] + P[(0)(2,1,3)] + P[(0)(3,1,2)] + P[(2)(1,3)]$$

$$+ P[(1)(2,3)] + P[(2)(3,1)] + P[(1)(3,2)] + P[(1,2)(3)]$$

$\rho_{11} = \rho_{12} = \rho_{13} = 0$ (by definition)

$$\rho_{2,1} = P[(0)(1,2,3)] + P[(3)(1,2)] + P[(2)(1,3)] + P[(2,3)(1)]$$

$$\rho_{2,2} = P[(0)(2,1,3)] + P[(3)(2,1)] + P[(1)(2,3)] + P[(1,3)(2)]$$

$$\rho_{2,3} = P[(0)(3,1,2)] + P[(2)(3,1)] + P[(1)(3,2)] + P[(1,2)(3)]$$

$$T_{1,1} = T_{2,1} = \rho_{2,1}\ \mu_{2,1}$$

$$T_{1,2} = T_{2,2} = \rho_{2,2}\ \mu_{2,2}$$

$$T_{1,3} = T_{2,3} = \rho_{2,3}\ \mu_{2,3}$$

$$R_{1,1} = 1/\mu_{1,1} \qquad R_{2,1} = L_{2,1}/T_{2,1}$$

$$R_{1,2} = 1/\mu_{1,2} \qquad R_{2,2} = L_{2,2}/T_{2,2}$$

$$R_{1,3} = 1/\mu_{1,3} \qquad R_{2,3} = L_{2,3}/T_{2,3}\ .$$

Performance metrics of the individual service center are obtained via the following:

$$L_1 = L_{1,1} + L_{1,2} + L_{1,3}$$

$$L_2 = L_{2,1} + L_{2,2} + L_{2,3}$$

$$T_1 = T_2 = T_{1,1} + T_{1,2} + T_{1,3}$$

$$R_1 = (R_{1,1}\ T_{1,1} + R_{1,2}\ T_{1,2} + R_{1,3}\ T_{1,3})\ /\ T_1$$

$R_2 = (R_{2,1}\ T_{2,1} + R_{2,2}\ T_{2,2} + R_{2,3}\ T_{2,3})\ /\ T_2$

$\rho_1 = 0$

$\rho_2 = \rho_{2,1} + \rho_{2,2} + \rho_{3,2}$.

For example if

$$\mu_{11} = \mu_{12} = \mu_{13} = 500$$

$$\mu_{21} = \mu_{22} = \mu_{23} = 1{,}000 \ ,$$

then the steady-state probabilities are:

$P[(0)(1,2,3)] = 0.06203$

$P[(0)(2,1,3)] = 0.05468$

$P[(0)(3,1,2)] = 0.04119$

$P[(3)(1,2)] = 0.05754$

$P[(2)(1,3)] = 0.06653$

$P[(3)(2,1)] = 0.03340$

$P[(1)(2,3)] = 0.07536$

$P[(2)(3,1)] = 0.04119$

$P[(1)(3,2)] = 0.04119$

$P[(2,3)(1)] = 0.09023$

$P[(1,3)(2)] = 0.10199$

$P[(1,2)(3)] = 0.12357$

$P[(1,2,3)(0)] = 0.21053$

and the performance metrics are:

$L_{1,1} = 0.55263$	$L_{2,1} = 0.44737$
$L_{1,2} = 0.53204$	$L_{2,2} = 0.46796$
$L_{1,3} = 0.49428$	$L_{2,3} = 0.50572$
$L_1 = 1.57895$	$L_2 = 1.42105$
$T_{1,1} = 276.32$	$T_{2,1} = 276.32$
$T_{1,2} = 266.02$	$T_{2,2} = 266.02$
$T_{1,3} = 247.14$	$T_{2,3} = 247.14$
$T_1 = 789.48$	$T_2 = 789.48$
$R_{1,1} = 2E-3$	$R_{2,1} = 1.6190E-3$
$R_{1,2} = 2E-3$	$R_{2,2} = 1.7591E-3$
$R_{1,3} = 2E-3$	$R_{2,3} = 2.0463E-3$
$R_1 = 2.0E-3$	$R_2 = 1.8000E-3$
$\rho_{1,1} = 0$	$\rho_{2,1} = 0.27632$
$\rho_{1,2} = 0$	$\rho_{2,2} = 0.26602$
$\rho_{1,3} = 0$	$\rho_{2,3} = 0.24714$
$\rho_1 = 0$	$\rho_2 = 0.78948$.

Some additional comments are appropriate. First, the procedure cannot be applied to all closed queueing networks. The network must be be representable by a 'pure' Markov process. That is, all of the service time distributions must be exponential or must be able to be represented by exponential stages. Note that if the service time distribution contains a discontinuity, then an infinite number of exponential stages is required. For example, if service time is deterministic, then it cannot be represented by the method of stages. There is also a practical limit to the number of equations that can be solved. Recall that the number of network states increases rapidly with the number of customers, chains and service centers. It also increases rapidly if the service time must be represented by the method of stages. It is not unusual for even a small two service center network to have well over 1,000 states. Finally, as previously stated, an open network that limits the customer population to some finite number is equivalent to a closed network. Thus, in theory the procedure can be applied to these networks, however, more often than not the number of feasible network states prohibits it.

CHAPTER 8

CONCLUSIONS

8.1 The Deceptive Service Center

Figure 8.1 depicts a simple service center. Customers arrive at the service center, wait in line for one of the two servers to become free, receive service, and depart. One would surely think that for such a simple system equations for the mean performance metrics could be derived. However, this is an open research problem, and has only been solved for two special cases [LAVE83]. These cases were discussed in this text and are (1) the M/M/m system, and (2) the M/G/m system with LCFSPR service discipline.

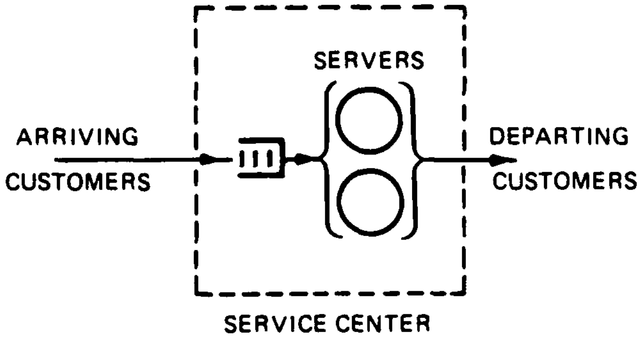


Figure 8.1 A 'Simple' Service Center.

Obviously, this 'simple' problem has been around for quite some time, and what makes it so deceptive is its simple representation. This problem symbolizes the paradoxical nature of the discipline.

8.2 Foundation

Most of the material covered in chapters 1 through 4 contains material normally taught in a graduate level course(s). The material was presented here as foundation to support the thrust of this text: 'Markovian Network Theory'. Clearly, before one can analyze a network of service centers, one must first learn how to analyze single service centers.

8.3 Contributions

Chapter 5 contains the only known example of local balance being used to solve a network of two or more service centers. In addition, it was illustrated by examples that if the arrival rate to a network varied according to the number of customers in the system (up to some finite number after which arrival ceased), then the network could be mapped into an equivalent closed network (also applies to networks in Chapter 6). This was stated by other authors, but again, these are believed to be the only known examples in the literature. It was also stated by others that the service rate of a service center could be a function of both the number of customers in the service center and of the number of customers in a subset of service centers, however, no references were given. This was proved in Chapter 5.

Credit for the pioneering work in Chapter 6 belongs to the authors Baskett, Chandy, Muntz, and Palacios, however, in order to clear up the ambiguities in their paper it was necessary to rederive most of their equations and to derive some that are not present. The reasons for this were : (1) When representing nonexponential service time by the method of exponential stages there is a finite probability that a customer will experience a zero length service time. They acknowledged this in their paper, but they did not include its existence in their derivation. Such an existence was illustrated in the example given in section 6.2 where this probability was 1/4. (2) Even for the less general case that they did consider, their equations for $f_i(x_i)$ which contain terms that account for the stages, are erroneous (indices and subscript errors).

As a result of considering the more general case, a general equation for the mean service time, in terms of the mean service time at each stage was derived. This equation is necessary in order to derive the equations for the aggregate network states, but such an equation was not given in the original paper. A third aggregate state, which combines all classes in a chain into one 'equivalent class', was not included in their paper, but followed easily from the rederivation (this state was alluded to by other authors and probably exists somewhere in the literature, but was not found). This state significantly reduces the amount of work that is required to determine the normalizing constant.

The process of rederiving the equations for the $f_i(x_i)$'s resulted in a much clearer understanding of types PS and IS service centers. There is never a queue or waiting line at these types of service centers. In addition, all customers are receiving service simultaneously. Thus, for type IS service centers each stage behaves as an independent service center. The only difference for type PS service centers is that the service rate at each stage depends upon the total number of customers. It was proved in Chapter 5 that the service rate of a subset of service centers could depend on the number of customers in the subset. Thus, in both cases the stages behave as service centers with an exponential service time.

It follows from this that the proofs given in Chapter 5 are sufficient for networks with types FCFS, IS, and PS service centers (classes and nonexponential service times are allowed at IS and PS service centers, but not at FCFS service centers). The theorem given in the paper applies to these cases and to LCFSPR service centers with nonexponential service times and classes. In addition classes are allowed at FCFS service centers under the constraint that they have the same exponential service time distribution function. It is true, as the authors state, that if the local balance equations are satisfied (apply), then the theorem holds. However, the difficulty is in proving that the local balance applies for any network composed of these types of service centers. It can be shown to apply for specific networks, but this was not done in their paper. There are several examples in this text showing that local balance applies.

There are also several examples showing that if the network contains service centers other than these types then local balance is not applicable. For example, if the service discipline is FCFS and different mean service times are allowed for different classes, then the local balance equations are inconsistent. Another counter example is when the service discipline is nonpreemptive priority, in this case not all the global balance equations can be subdivided into local balance equations, hence local balance does not apply.

Chapter 7 contains numerous examples illustrating how to apply the Mean Value Analysis (MVA) algorithm to closed networks. Examples for the following types of networks were given: single chain, load independent; single chain, load dependent; multiple chain, load independent; multiple chain, load dependent; and single closed chain, load independent mixed network. It was stated in this chapter that these were the only known examples of the MVA algorithm, however, some have been found for cyclic networks (cyclic networks are a special case of single chain, load independent networks).

In addition the concept of using the throughput theorem in conjunction with MVA to obtain the normalizing constant is not presented elsewhere. At present, there is no algorithm that can always prevent overflow when trying to determine the normalizing constant [LAVE83]. In the event of overflow these algorithms fail. This is irrelevant when using MVA since it does not depend on the determination of this constant in order to determine the mean performance metrics. The advantage of using the throughput theorem with MVA is that when overflow is not a problem the normalizing constant can be obtained and even if overflow does occur, the mean performance metrics can still be obtained. The disadvantage of this technique is that it requires more memory, however, this is often not a problem.

Some of the contributions claimed may quite possibly appear somewhere in the literature (in particular, the aggregate state that deals with chains, and the stage equation for the mean service time), but the fact that they were not found and had to be rederived indicates that the material badly needed unifying. This was accomplished in this work, and because of the subject matter and volume of the material, it is believed to be a major contribution.

8.4 A Characterization of Networks with Product Form Solutions

For a network of service centers to have a product form solution each service center in the network must meet one of the following sets of conditions [CHAN77] [CHAN83]:

(1) If the service discipline is FCFS, then the mean service time of all customer classes must be the same, and the distribution must be negative exponential (each customer class may have its own set of routing probabilities).

(2) If the service discipline is such that every customers starts to receive some service immediately upon arriving, then each customer class may have its own general service time distribution (the density function must have a rational Laplace transform) and routing probabilities. Note, service center types LCFSPR, PS and IS meet the condition that each customers starts to receive some service immediately upon arrival.

In addition, if the network is open, then all arrival processes from outside the network must be Poisson, and no queue can saturate (utilization > 1).

If a network meets these conditions then the steady-state probability that the network is in state $(x_1,x_2,...,x_N)$ is given by:

$$P(x_1,x_2,...,x_N) = \frac{P_1(x_1)\ P_2(x_2)\ \ldots\ P_N(x_N)}{G} ,$$

where N equals the number of service centers, x_i represents the conditions prevailing at service center i, $P_i(x_i)$ is a factor corresponding to probability that service center i is in state x_i, and G is a normalizing constant chosen to make the probabilities sum to one.

The factor $P_i(x_i)$ contains only parameters that pertain to service center i. It is the same factor that results from assuming that the arrival process is Poisson and analyzing the service center in isolation. If the network is open and the arrival process does not depend on the number of customers, then the mean arrival rate can be uniquely determined. However, if the network is closed, then the arrival rate (same as relative throughput) can only be determined relative to the arrival rates at the other service centers. In this case a positive value is assigned to the arrival rate at one of the service centers. The others can then be determined (the normalizing constant will compensate for this). What is amazing about this is that, in general, the arrival process at the individual service centers are not Poisson! What is even more amazing is that if one knew exactly what the arrival process was he could not obtain an exact solution, since at present only partial results have been obtained for the G/M/1 system.

It is important to emphasize that all of the solutions thus far have been obtained by guessing at the answer or by local balance (a form of guessing). In other words, the solutions can be proved by substituting the guessed at results into the global balance equations and determining if they are satisfied. However, at present, there is no way to derive the results.

8.5 Approximate Solution to Queueing Networks

If a network does not have a product form solution then the distribution of customers at a service center is a function not only of the service center in question, but also of other parameters in the network. This is an extremely difficult problem. One approach is to make an assumption that allows an answer to be obtained (usually the assumption is true for networks with product form solutions and hopefully an approximation for other networks), then to validate the assumption by showing that the answer is close to the solution to the problem that was (presumably) unsolvable. In other words an answer has to be obtained by some other means than queueing theory such as simulation. If the assumption is tested for similar networks and it also holds (produces small errors), then one can use it in the future on similar networks without testing it. Such solutions are heuristic and cannot be formally defended. At present all of the techniques for obtaining approximate solutions fall into this category [SAUE81]. Hence, for this reason they were not covered in this text. There are quite a few heuristic approaches that can be used on closed networks, but few have been extended or shown to hold for open networks (some have been shown not to be applicable [LAVE83]). One of the reasons for this is that closed networks are in a sense self regulating and therefore, more predictable. Even for closed networks without product-form solutions bounds can often be obtained, although they are somewhat loose. It should be obvious that the development of heuristic techniques is a trial and error procedure. In addition, one can spend considerably more time trying to validate them than developing them.

8.6 Review of Latest Textbooks

Although there is still not a comprehensive text devoted to the area of queueing network theory, three texts that use the theory were published during the course of this research.

1. A Computer and Communications Network Performance Analysis, by B.W. Stuck and E. Arthurs, Printice-Hall, 1985.

2. Performance Analysis of Local Computer Networks, by J.L. Hammond and J.P. O'Reilly, Addison-Wesley, 1986.

3. Telecommunications Network : Protocols, Modeling and Analysis, M. Schwartz, Addisson-Wesley, 1987.

8.6.1 Textbook by Stuck and Arthurs

The text by Stuck and Arthurs is the extreme opposite in almost every way of the material covered here. As they indicate in their preface, virtually no time at all is spent deriving results. Their philosophy is suggested in the preface as: 'The crux of engineering, in our opinion, is manipulating numbers in a great variety of ways to gain qualitative insight into design issues via quantitative methods'.

A second motivation is given in terms of the mind-set of their Bell Laboratory graduate engineers: 'Many of them are simply not interested in derivations'. Therefore, the widely used approach for deriving fundamental results was purposely ignored in this text, whereas in contrast, it is the derivation of fundamental relationships that constitutes the purpose and content of this dissertation.

The book primarily focuses on closed networks and obtaining bounds for such networks. Jackson-type networks are introduced in their Chapter 6. The following two chapters are concerned with applications of Jackson-type networks. However, the emphasis is heavily on obtaining bounds for these types of networks. Several hours were spent reviewing these two chapters, and although some Jackson-type equations appear, it is doubtful that there is a single problem worked out using them. The material on Jackson-type networks is perhaps the worst in the text. There is some material on open networks, however it is presented in such a way that it is extremely difficult to extricate it from that on closed networks. In order to get around having to explain how to determine the normalizing constant, a computer program is simply given. The approach use in these chapters seems to be to cover as many cases as possible and provide equations for these, rather than to stress the fundamentals and provide a few general examples.

Single service centers are discussed, but only after networks of service centers. Again quoting from the preface, the material in these chapters (the last two) is the most mathematically sophisticated, and requires the greatest intellectual maturity. It should also be noted that the only systems discussed in them are of the type M/G/1. The M/M/1 and M/M/m systems are not discussed here or anywhere else in the book.

8.6.2 Textbook by Hammond and O'Reilly

Hammond and O'Reilly devote very little effort to deriving the basic queueing theory equations. For example, the mean performance metrics of the M/G/1 model are developed first. The equations for the M/M/1 are developed from these. In the chapter on queueing theory a total of three pages are spent on queueing networks. The emphasis is on explaining the different types of queueing networks: open, closed and mixed. The discussion on mixed network is ambiguous since the figure referenced is an open network with feedback. Also, depicted in this section is an open network with three service centers in tandem (the output of the first feeds directly to the input of the second, etc.) and it is pointed out that: 'this type of queueing network is of the type most often used for modeling multiaccess networks, and thus attention is restricted to this type'. Burke's Theorem is then quoted and it is stated that this type of network can be broken down into a collection of M/M/1 and M/G/1 submodels. However, it should be noted that this statement cannot be formally defended.

Burke's Theorem states that the output process of a M/M/m system is Poisson and independent of all other processes in the system. However, Burke also proved that it is the only such FCFS system with this property [KLEI75]. The procedure is used in the chapter on ring networks, and it is believed that their justification is contained in the following two sentences: 'The interarrival times and the service times are independent, thus the chain of station elements in Fig 8.22 can be broken up into independent submodels for each station. This assumption, which results in a very tractable model, has been shown by comparison to simulation results to give reasonably accurate, although somewhat pessimistic results'. While the statement that the interarrival times are independent is true, the arrival process is the sum of two processes, one of which is Poisson and the other non-Poisson. Thus, it is believed that the second sentence is actually the justification for the first.

The following is quoted from the cover of the book 'the performance models discussed are developed in as elementary a manner as possible, often using a heuristic rather than a rigorous approach'. There is nothing wrong with the approach, and it is one of the strengths of the book, yet, it is necessary to know when assumptions such as these can be made. If it is not clearly stated as an assumption, and the conditions under which it is applicable are not pointed out, then one is likely to misuse it.

8.6.3 Textbook by Schwartz

The book by schwartz relates more than any of the others to the work contained in this dissertation. He devotes an entire chapter to developing the fundamental queueing theory equations. However, compared to Chapter 2,3 and 4 of this work the book is somewhat brief.

For example, the strong connection between Markov processes and queueing theory is not explored. The M/M/1 and M/M/2 are analyzed as birth and death processes, however there is no mention of fact that this is a special case of a Markovian process. He brings out the fact that in the M/G/1 system the service time is not memoryless, and one can no longer set up a simple balance equation for states of the system. He then proceeds to look at the system at only departure instants without an explanation as to why this is being done (i.e., because an embedded Markov process exists at these points). He then uses the same trick (his terminology), as in Hammond and O'Reilly, to arrive at the mean performance equations without going through the detailed analysis that was given in Chapter 4 of this work.

There is even a section on queueing network theory in Chapter 5 of the book. Again it is brief, but probably adequate for application purposes. For example, he does state that a queueing network is a multidimenional birth and death process and writes out the general equations for an open queueing network with exponential service times. He does not develop the equations for closed networks, but simply states them. He does discuss techniques for determining the normalizing constant, but only considers single chain networks with load independent service rates. From an applications point of view the material in this section of the book is close to the material in Chapters 5 (this work). However, there are salient differences. For example Burke's Theorem, local balance, and the fact the arrival processes are not Poisson, are not even mentioned. In addition there are a few partial examples, but none worked from beginning to end as those in Chapter 5 of this work.

It is not surprising that service disciplines such as PS and LCFSPR are not covered in this text since it is on communications systems. However, it is somewhat surprising that classes and type IS service centers are not covered in the book. As demonstrated by the example in Chapter 6 of this work, classes are sometimes necessary in communications system in order to describe the routing of messages. In addition IS service centers can be used to account for propagation and other delays that are encountered in networks that stretch across the nation. In contrast to the text by Hammond and O'Reilly, this author is quite rigorous when it comes to specifying that a technique is an approximation or that an assumption is made to get answer. In many cases he does specify the range over which the approximation or assumption is valid.

8.6.4 Final Remarks

One of the common failings in each of the three books reviewed is that they all delegate very little space to the derivation of basic queueing theory equations. The material is conceptually complex and can be mathematically intimidating to such an extent that authors feel that it is necessary to gloss over it so as to have room to present their applications techniques. Understandably, since the texts are applications oriented, one would expect them to be heavily weighted in this direction. It should be brought out that this gap needs to be bridged with references that are in themselves understandable.

It was one of the purpose of the dissertation to enhance that understandability and thereby narrow that gap. It has done so by going to the source material developed by pioneers in the discipline. Where the explanations have been brief and obscure, they have been expanded and clarified. Where illustrative examples where not given, they were worked out. Where proofs could not be found, they were developed.

APPENDIX A

SUPPLEMENTARY LOCAL BALANCE PROBLEMS

A.1 FCFS with Two Classes

Figure A.1 depicts a queueing network composed of one service center with two customer classes. It is assumed the customers arrive from Poisson sources. The mean arrival rates of class 1 and class 2 customers is λ_1 and λ_2 respectively. It is also assumed that the service time distributions of both classes are exponentially distributed. However, the mean service rates may or may not be the same. The network will be analyzed for both FCFS and LCFSPR service disciplines.

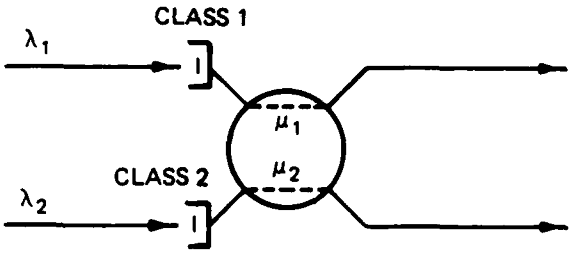


Figure A.1 Service Center with Two Customer Classes.

The following analysis assumes that the service discipline is FCFS and that the rates μ_1 and μ_2 are different. The global balance equations for states with two or fewer customers are:

$$(\lambda_1+\lambda_2)P(0) = \mu_1P(1) + \mu_2P(2)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1) = \mu_1P(1,1) + \mu_2P(2,1) + \lambda_1P(0)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2) = \mu_1P(1,2) + \mu_2P(2,2) + \lambda_2P(0)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,1) = \mu_1P(1,1,1) + \mu_2P(2,1,1) + \lambda_1P(1)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,2) = \mu_1P(1,1,2) + \mu_2P(2,1,2) + \lambda_2P(1)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,1) = \mu_1P(1,2,1) + \mu_2P(2,2,1) + \lambda_1P(2)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,2) = \mu_1P(1,2,2) + \mu_2P(2,2,2) + \lambda_2P(2) \ .$$

Observe that there are only seven equations and 15 unknowns. Also, note that no matter how many equations are written there will always be more unknowns than equations.

Now assume that local balance holds. That is, the rate of flow out of a state due to customer of class c departing is equal to the rate of flow into the same state due to the arrival of a class c customer. The local balance equations that correspond (the sum of the local balance equations are the global balance equations) to the seven global equations are :

$$\lambda_1P(0) = \mu_1P(1)$$

$$\lambda_2P(0) = \mu_2P(2)$$

$$\lambda_1P(1) = \mu_1P(1,1)$$

$$\lambda_2P(1) = \mu_2P(2,1)$$

$$\mu_1P(1) = \lambda_1P(0)$$

$$\lambda_1P(2) = \mu_1P(1,2)$$

$$\lambda_2P(2) = \mu_2P(2,2)$$

$$\mu_2P(2) = \lambda_2P(0)$$

$$\lambda_1 P(1,1) = \mu_1 P(1,1,1)$$

$$\lambda_2 P(1,1) = \mu_2 P(2,1,1)$$

$$\mu_1 P(1,1) = \lambda_1 P(1)$$

$$\lambda_1 P(1,2) = \mu_1 P(1,1,2)$$

$$\lambda_2 P(1,2) = \mu_2 P(2,1,2)$$

$$\mu_1 P(1,2) = \lambda_2 P(1)$$

$$\lambda_1 P(2,1) = \mu_1 P(1,2,1)$$

$$\lambda_2 P(2,1) = \mu_2 P(2,2,1)$$

$$\mu_2 P(2,1) = \lambda_1 P(2)$$

$$\lambda_1 P(2,2) = \mu_1 P(1,2,2)$$

$$\lambda_2 P(2,2) = \mu_2 P(2,2,2)$$

$$\mu_2 P(2,2) = \lambda_2 P(2) \ .$$

The following is a subset of the local balance equations :

$$\mu_1 P(1) = \lambda_1 P(0)$$

$$\mu_2 P(2) = \lambda_2 P(0)$$

$$\mu_1 P(1,1) = \lambda_1 P(1)$$

$$\mu_1 P(1,2) = \lambda_1 P(2)$$

$$\mu_1 P(1,2) = \lambda_2 P(1)$$

$$\mu_2 P(2,1) = \lambda_2 P(1)$$

$$\mu_2 P(2,1) = \lambda_1 P(2)$$

$$\mu_2 P(2,2) = \lambda_2 P(2) \ .$$

Solving these equations in terms of P(0) results in :

$$P(1) = (\lambda_1)\ (1/\mu_1)\ P(0)$$

$$P(2) = (\lambda_2)\ (1/\mu_2)\ P(0)$$

$$P(1,1) = (\lambda_1)^2\ (1/\mu_1)^2\ P(0)$$

$$P(1,2) = (\lambda_1)\ (\lambda_2)\ (1/\mu_1)\ (1/\mu_2)\ P(0)$$

$$P(1,2) = (\lambda_1)\ (\lambda_2)\ (1/\mu_1)^2\ P(0)$$

$$P(2,1) = (\lambda_1)\ (\lambda_2)\ (1/\mu_1)\ (1/\mu_2)\ P(0)$$

$$P(2,1) = (\lambda_1)\ (\lambda_2)\ (1/\mu_2)^2\ P(0)$$

$$P(2,2) = (\lambda_2)^2\ (1/\mu_2)^2\ P(0)\ .$$

Observe that there are two equations for P(1,2) and P(2,1). Also notice that they are inconsistent. The conclusion is that local balance does not apply if the service rates are different for the two classes. However, if $\mu_1=\mu_2=\mu$, then local balance does hold, and the form of the solution is :

$$P(x_1,x_2,\dots,x_k) = [\lambda_1^{k_1}\ \lambda_2^{k_2}]\ [(1/\mu)^{k_1+k_2}]\ P(0)$$

where k_1 and k_2 equals the number of class 1 and class 2 customers respectively.

A.2 LCFSPR with Two Classes

If the service discipline of the network in Figure A.1 is changed to LCFSPR, then the global balance equations are:

$$(\lambda_1+\lambda_2)P(0) = \mu_1 P(1) + \mu_2 P(2)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1) = \mu_1 P(1,1) + \mu_2 P(2,1) + \lambda_1 P(0)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2) = \mu_1 P(1,2) + \mu_2 P(2,2) + \lambda_2 P(0)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,1) = \mu_1 P(1,1,1) + \mu_2 P(2,1,1) + \lambda_1 P(1)$$

$$(\lambda_1+\lambda_2+\mu_1)P(1,2) = \mu_1P(1,1,2) + \mu_2P(2,1,2) + \lambda_1P(2)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,1) = \mu_1P(1,2,1) + \mu_2P(2,2,1) + \lambda_2P(1)$$

$$(\lambda_1+\lambda_2+\mu_2)P(2,2) = \mu_1P(1,2,2) + \mu_2P(2,2,2) + \lambda_2P(2) .$$

The corresponding local balance equations are:

$$\lambda_1P(0) = \mu_1P(1)$$

$$\lambda_2P(0) = \mu_2P(2)$$

$$\lambda_1P(1) = \mu_1P(1,1)$$

$$\lambda_2P(1) = \mu_2P(2,1)$$

$$\mu_1P(1) = \lambda_1P(0)$$

$$\lambda_1P(2) = \mu_1P(1,2)$$

$$\lambda_2P(2) = \mu_2P(2,2)$$

$$\mu_2P(2) = \lambda_2P(0)$$

$$\lambda_1P(1,1) = \mu_1P(1,1,1)$$

$$\lambda_2P(1,1) = \mu_2P(2,1,1)$$

$$\mu_1P(1,1) = \lambda_1P(1)$$

$$\lambda_1P(1,2) = \mu_1P(1,1,2)$$

$$\lambda_2P(1,2) = \mu_2P(2,1,2)$$

$$\mu_1P(1,2) = \lambda_1P(2)$$

$$\lambda_1P(2,1) = \mu_1P(1,2,1)$$

$$\lambda_2P(2,1) = \mu_2P(2,2,1)$$

$$\mu_2P(2,1) = \lambda_2P(1)$$

$$\lambda_1 P(2,2) = \mu_1 P(1,2,2)$$

$$\lambda_2 P(2,2) = \mu_2 P(2,2,2)$$

$$\mu_2 P(2,2) = \lambda_2 P(2) \ .$$

The following is a subset of the local balance equations :

$$\mu_1 P(1) = \lambda_1 P(0)$$

$$\mu_2 P(2) = \lambda_2 P(0)$$

$$\mu_1 P(1,1) = \lambda_1 P(1)$$

$$\mu_1 P(1,2) = \lambda_1 P(2)$$

$$\mu_2 P(2,1) = \lambda_2 P(1)$$

$$\mu_2 P(2,2) = \lambda_2 P(2) \ .$$

There are six equations and seven unknowns. Solving these in terms of P(0) results in :

$$P(1) = (\lambda_1)\ (1/\mu_1)\ P(0)$$

$$P(2) = (\lambda_2)\ (1/\mu_2)\ P(0)$$

$$P(1,1) = (\lambda_1)^2\ (1/\mu_1)^2\ P(0)$$

$$P(1,2) = (\lambda_1)\ (\lambda_2)\ (1/\mu_1)\ (1/\mu_2)\ P(0)$$

$$P(2,1) = (\lambda_1)\ (\lambda_2)\ (1/\mu_1)\ (1/\mu_2)\ P(0)$$

$$P(2,2) = (\lambda_2)^2\ (1/\mu_2)^2\ P(0) \ .$$

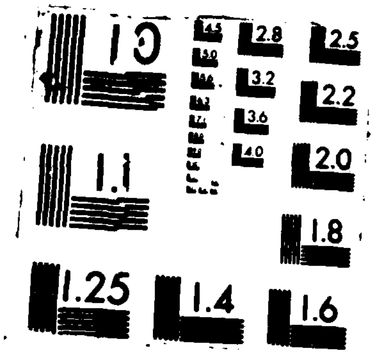
The form of the solution is:

$$P(x_1,x_2,\ldots,x_k) = [\lambda_1^{k_1}\ \lambda_2^{k_2}]\ [(1/\mu_1)^{k_1}\ (1/\mu_2^{k_2}]\ P(0)$$

where k_1 and k_2 equals the number of class 1 and class 2 customers respectively.

AD-A182 994 NETWORKS OF MARKOVIAN QUEUES(U) ARMY ENGINEER WATERWAYS 4/4
EXPERIMENT STATION VICKSBURG MS INSTRUMENTATION
SERVICES DIV R A FRANCO MAY 87 WES/TR/O-87-2

UNCLASSIFIED F/G 12/3 NL


1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

A.3 LCFSPR Analysis with Two Exponential Stages

A queueing network that consists of a single service center is depicted in Figure A.2. There is only one customer class, but the service time is represented by two exponential stages, and the service discipline is last-come-first-serve-preemptive resume.

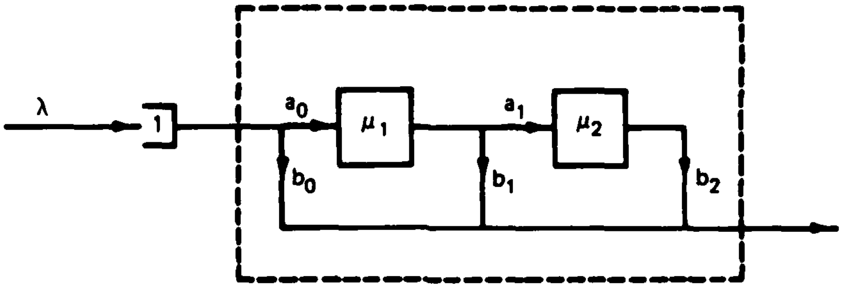


Figure A.2 Service Center with Two Exponential Stages.

In order that the state summarize all past history of the process, it must contain the stage that each customer was in when preempted and his order at the service center. If the service center contains k customers, then let $(m_1,m_2,\ldots,m_k)$ be the state, where m_1 is the stage of the last customer to arrive.

By Equating the rate of flow into and out of a state the following global balance equations, for states in which there are two or fewer customers, are obtained :

$a_0\lambda P(0) = \mu_1 b_1 P(1) + \mu_2 P(2)$

$(\mu_1+a_0\lambda)P(1) = a_0\lambda P(0) + \mu_1 b_1 P(1,1) + \mu_2 P(2,1)$

$(\mu_2+a_0\lambda)P(2) = \mu_1 a_1 P(1) + \mu_1 b_1 P(1,2) + \mu_2 P(2,2)$

$(\mu_1+a_0\lambda)P(1,1) = a_0\lambda P(1) + \mu_1 b_1 P(1,1,1) + \mu_2 P(2,1,1)$

$$(\mu_1+a_0\lambda)P(1,2) = a_0\lambda P(2) + \mu_1 b_1 P(1,1,2) + \mu_2 P(2,1,2)$$

$$(\mu_2+a_0\lambda)P(2,1) = \mu_1 a_1 P(1,1) + \mu_1 b_1 P(1,2,1) + \mu_2 P(2,2,1)$$

$$(\mu_2+a_0\lambda)P(2,2) = \mu_1 a_1 P(1,2) + \mu_1 b_1 P(1,2,2) + \mu_2 P(2,2,2) .$$

In the first three equations there are seven unknowns. In the first seven equations there are fifteen unknowns. Again, no mater how many equations are written there will always be more unknowns than equations.

The corresponding local balance equations are obtained by equating the rate of flow out of a state due a customer leaving a stage of service to the rate of flow into that state due to customer entering that stage:

$$a_0\lambda P(0) = \mu_1 b_1 P(1) + \mu_2 P(2)$$

$$\mu_1 P(1) = a_0\lambda P(0)$$

$$a_0\lambda P(1) = \mu_1 b_1 P(1,1) + \mu_2 P(2,1)$$

$$\mu_2 P(2) = \mu_1 a_1 P(1)$$

$$a_0\lambda P(2) = \mu_1 b_1 P(1,2) + \mu_2 P(2,2)$$

$$\mu_1 P(1,1) = a_0\lambda P(1)$$

$$a_0\lambda P(1,1) = \mu_1 b_1 P(1,1,1) + \mu_2 P(2,1,1)$$

$$\mu_1 P(1,2) = a_0\lambda P(2)$$

$$a_0\lambda P(1,2) = \mu_1 b_1 P(1,1,1) + \mu_2 P(2,1,2)$$

$$\mu_2 P(2,1) = \mu_1 a_1 P(1,1)$$

$$a_0\lambda P(2,1) = \mu_1 b_1 P(1,2,1) + \mu_2 \ P(2,2.1)$$

$$\mu_2 P(2,2) = \mu_1 a_1 P(1,2)$$

$$a_0 \lambda P(2,2) = \mu_1 b_1 P(1,2,2) + \mu_2 P(2,2,2) \ .$$

The sum of the local balance equations are the global balance equations. Therefore the solution of the local balance equations will satisfy the global balance equations. The following is a subset of the local balance equations :

$$\mu_1 P(1) = a_0 \lambda P(0)$$

$$\mu_2 P(2) = \mu_1 a_1 P(1)$$

$$\mu_1 P(1,1) = a_0 \lambda P(1)$$

$$\mu_1 P(1,2) = a_0 \lambda P(2)$$

$$\mu_2 P(2,1) = \mu_1 a_1 P(1,1)$$

$$\mu_2 P(2,2) = \mu_1 a_1 P(1,2) \ .$$

Notice that there are six equations and seven unknowns. Solving these equations in terms of P(0) results in :

$$P(1) = \frac{a_0 \lambda}{\mu_1} P(0)$$

$$P(2) = \frac{a_0 a_1 \lambda}{\mu_2} P(0)$$

$$P(1,1) = \frac{(a_0 \lambda)^2}{(\mu_1)^2} P(0)$$

$$P(1,2) = \frac{a_0{}^2 a_1 \lambda^2}{\mu_1 \mu_2} P(0)$$

$$P(2,1) = \frac{a_0{}^2 a_1 \lambda^2}{\mu_1 \mu_2} P(0)$$

$$P(2,2) = \frac{(a_0 a_1 \lambda)^2}{(\mu_2)^2} P(0) .$$

The form of the solution for these six states is

$$P(m_1,m_2,\ldots,m_k) = \frac{f_1(m_1)\ f_2(m_2)\ldots f_k(m_k)}{G}$$

where $$f_i(m_i) = \begin{cases} a_0\lambda/\mu_1 & \text{for } m_i=1 \\ a_0 a_1 \lambda/\mu_2 & \text{for } m_i=2 \end{cases}$$

and $$G = 1/P(0) .$$

It is easy to verify that this is the solution for any state. It satisfies both the local and global balance equations. The solution also agrees with the equations in Chapter 6.

APPENDIX B

SOURCE LISTING OF MVA PROGRAM

FOR

MULTIPLE CHAIN, LOAD INDEPENDENT NETWORKS

```
Program MVA (input,output);

Const
  Max_M = 3;          {Maximum number of Service Centers}
  Max_C = 10;         {Maximum number of Chains or Classes}
  MaxSize = 1000;

Type
  Population_Vector = Array[1..Max_C] of Integer;
  Matrix = Array[1..Max_M,1..Max_C] of Real;
  ServerType = 0..1; {0 = IS , 1 = all others}

Var
  N,N_Max     : Population_Vector;
  L,T,R,V,S   : Matrix;
  Length      : Array[1..Max_M,0..MaxSize] of Real;
  Queue       : Array[1..Max_M] of ServerType;
  M           : integer;  {Number of Service Centers}
  C           : integer;  {Number of Chains or Classes}
  I,K         : Integer;
  X           : Real;

{----------------------------------------------------------------
This function increments the present population vector, N, and
is set to false, if N does not equal the maximum population
vector, N_Max. In addition to the above parameters the length
of both population vector, C, is passed to the function.        }

Function Increment_Population_Vector
                                  (Var N : Population_Vector;
                                   N_Max : Population_Vector;
                                       C : Integer): Boolean;
Var
  Flag : Boolean;
  J    : Integer;
```

```
Begin
  J := 1;  {always start from the far left and proceed to the right}
  Repeat
    N[J] := N[J] + 1;
    If N[J] <= N_Max[J] then Flag := True
    Else   {reset present column and increment next column}
      Begin
        Flag := False;
        N[J] := 0;
        J := J + 1
      End;
  Until ((J > C) Or Flag);
  Increment_Population_Vector := (J > C)
End;

{-----------------------------------------------------------------
Given a population vector, N, this function computes the row
index value (0..MaxSize) of the matrix Length. The formula is :
Index := n1 + [N_Max(1)+1] n2 + [N_Max(1)+1][N_Max(2)+1] n3 +
         ... + [N_Max(1)+1][N_Max(2)+1]...[N_Max(C-2)+1] n(C-1)
where n1,n2,...,nC are the elements of the population vector N,
and N_Max(1),N_Max(2),...,N_Max(C) are the elements of the maximum
population vector,N_Max. Note the maximum index value is :
    [N_max(1)+1][N_Max(2)+1]...[N_Max(C-1)], and is not
    [N_max(1)+1][N_Max(2)+1]...[N_Max(C-1)][N_Max(C)].
Hence a considerable memory saving is accomplished by writing over
Length values that are no longer required by the MVA algorithm.
To take maximum advange of this saving the chain with the largest
population should be nC.                                         }

Function Index (N,N_Max : Population_Vector; C:Integer): Integer;
Var
  J,Sum,Radix : Integer;
Begin
  If C = 1 then Index := 0
  Else
    Begin
      Sum := N[1];
      Radix := N_Max[1] + 1;
      For J := 2 To (C-1) do
      Begin
        Sum := Sum + ( N[J] * Radix );
        Radix := Radix * ( N_Max[J] + 1 )
      End; (* for *)
      Index := Sum
    End (* else *)
End; (* Index *)
```

```
{-----------------------------------------------------------------}
Begin (* main program *)
Writeln;
(* read parameters specific to this model *)
  write('Number of Service Centers   => '); Readln(M);
  For I := 1 to M do
  begin
    write('Service Center ',I:2,'  is type  => ');
    readln(Queue[I]);
  end;
  write('Number of Chains or Classes => '); Readln(C);
  For K := 1 to C do
  begin
    write('Number of jobs in chain ',K:2,'  => ');
    readln(N_Max[K]);
  end;
  writeln;
  For I:=1 to M do
    For K:= 1 to C do
    begin
      write('Relative number of visits [Center ',I:2,
      ', Chain ',K:2,'] => ');
      readln(V[I,K])
    end;
  writeln;
  For I:=1 to M do
    For K:=1 to C do
    begin
      write('Mean Service Time [Center ',I:2,' , Chain ',K:2, '] => ');
      readln(S[I,K])
    end;

  {initial parameters}
   For I := 1 to M do Length[I,0] := 0.0;
   (* initialize N = [1..C] := 0 *)

   For K := 1 to C do N[K] := 0;
   N[1] := 1;  {initalize population vector N := [1,0,0,...,0]}


  (* Perform calulations *)
  Repeat
    For K := 1 to C do
      Begin
        If N[K] = 0 then (* there are 0 jobs in chain k *)
          For I := 1 to M do L[I,K] := 0
        Else
         Begin (* caluate R *)
```

```
           For I := 1 to M do
             If Queue[I] = 0  (* if infinite server *)
               then R[I,K] := S[I,K]
             Else
               Begin
                 N[K] := N[K] - 1;
                 R[I,K] := S[I,K] * ( 1 + Length[I,Index(N,N_Max,C)] );
                 N[K] := N[K] + 1
               End;

           (* calulate throughput T *)
           (* calualate L by Class *)
           X := 0;
           For I := 1 to M do X:= X + V[I,K] * R[I,K];
           For I := 1 to M do
             Begin
               T[I,K] := N[K] * V[I,K] / X;
               L[I,K] := T[I,K] * R[I,K]
             End
         End (* else *)
      End; (* for *)
      (* calulate total L of each queue *)
      For I := 1 to M do
        Begin
          X := 0;
          For K := 1 to C do X := X + L[I,K];
          Length[I,Index(N,N_Max,C)] := X
        End;
    Until Increment_Population_Vector(N,N_Max,C);

  (* print performance parameters *)
  Writeln;
  Writeln;
  Writeln ('Center Chain   Throughtput    Response Time    Q-Length ');
  For i:=1 to M do
    For K:=1 to C do
    begin
      Write(' ',I:2);
      Write('     ',K:2);
      Write('     ',T[I,K]:10);
      Write('     ',R[I,K]:10);
      Writeln('        ', L[I,k]:10)
    end
End.
```

REFERENCES AND BIBLIOGRAPHY

ALLE78 Allen, O.A. Probability, Statisitcis, and Queuing Theory with Computer Science Applications. Academic Press, New York, 1978.

BART45 Bartlett, M.S. Proc. Cambridge Phil. Soc., 1945.

BASK75 Baskett, F., Chandy, K.M., Muntz, R.R., and Palacios, F. "Open, Closed, and Mixed Networks of Queues with Differerent Classes of Customers," J. ACM, April, 1975.

BASK72 Baskett, F., and Palacios, F. "Processor Sharing in a Central Server Queueing Model of Multiprogramming with Applications," Procc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton U., March, 1972.

BUZE73 Buzen, J.P. "Computational Algorithms for Closed Queuing Networks with Exponential Servers," Comm. ACM 16, September, 1976.

BRUE80 Bruell, S.C. Computational Algorithms for Closed Queueing Networks. Elsevier North Holland, New York, 1980.

CHAN72 Chandy, K.M., "The Analysis and Solutions for General Queueing Networks," Procc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton U., March, 1972.

CHAN77 Chandy, K.M., Howard, J.H., and Towsley, D.F. "Product Form and Local Balance in Queueing Networks," J. ACM, April, 1977.

CHAN78 Chandy, K.M., and Sauer, C.H. "Approximate Methos for Analysis Queueing Netork Models of Computing Systems, " ACM Computing Surveys, September, 1978.

CHAN80 Chandy, K.M., and Sauer, C.H. "Computation Algorithms for Product Form Networks," Communications of ACM, October, 1980.

CHAN83 Chandy, K.M., and Martin, A.J. "A Characterization of Product-Form Queueing Networks," J. ACM, April, 1983.

CLAR70 Clark, A.B., and Disney, R.L. Probability and Random Processes for Engineers and Scientists. Wiley-Interscience, New York, 1976.

COHE69 Cohen, J.W. The Single Server Queue. Elsevier North Holland, 1969.

COOP81 Cooper, R.B. *Introduction to Queuing Theory*, second ed. Elsevier North Holland, 1981.

COUR77 Courtois, P.J. *Decomposability - Queueing and Computer System Applications*. Academic Press, New York, 1977.

COX54 Cox. D.R. "A Use of Complex Probilities in the Theory of Stochastic Processes," *Proc. Cambridge Phil. Soc.*, 1955.

DENN78 Denning P.J., and Buzen J.P. "The Operational Analysis of Queueing Network Models," *ACM Computing Surveys*, September 1978.

FELL68 Feller, W. *An Introduction to Probability Therory and Its Applications*, Vol.1, 3rd ed. Wiley-Interscience, New York, 1968.

FRED67 Frederick S.H., and Lieberman G.J. *Introduction to Opetations Research*. Holden-Day, Inc, San Fanciso, 1967.

GORD67 Gordon, W.J., and Newell, G.J. "Closed Queueing Systems with Exponential Servers," *Operations Research 15*, March. 1967.

GROS74 Gross, D., and Harris, C.M. *Fundamentals of Queueing Theory*. Wiley-Interscience, New York, 1974.

HAYE84 Hayes J.F. *Modeling and Analysis of Computer Communications Networks*. Plenum Press, New York, 1984.

HAMM86 Hammond, J.L., and O'Reilly J.P. *Performance Analysis of Local Computer Networks*. Addison Wesely, Mass. 1986.

HAIG67 Haight F.A. *Handbook of the Poission Distribution*. Wiley-Interscience, New York, 1968.

JACK57 Jackson, J.R. "Networks of Waiting Lines," *Operations Research 5*, 1957

JACK63 Jackson, J.R. "Jobshop-like Queueing Systems," *Management Science 10*, October, 1963.

JAIS68 Jaiswal, N.K. *Priority Queues*. Academic Press, New York, 1968.

KELL75 Kelly, F.P. "Networks of Queues with Customers of Differenct Types," *J. Appl. Prob. 12*, 1975.

KLEI75 Kleinrock, L. *Queueing Systems, Volume 1: Theory*. Wiley-Interscience, New York, 1975.

KLEI76 Kleinrock, L. Queueing Systems, Volume 2: Computer Applications. Wiley-Interscience, New York, 1976.

KOBA77 Kobayashi, H., and Konheim, A.G. "Queueing Models for Computer Communications System Analysis," IEEE Trans. on Communications, January, 1977.

KOBA78 Kobayashi, H. Modeling and Analysis: An Introduction to Performance Evaluation Methodology. Addison-Wesely, Mass. 1978.

LAM83 Lam, S.S. "A Simple Derivation of the MVA and LBANC Algorithms from the Convolution Algorithm," IEEE Trans. Comput., November, 1983.

LAVE83 Lavenberge S.S. Computer Performance Modeling Handbook. Academic Press, New York, 1983.

LITT61 Little, J.D.C. "A Proof of the Queuing Formula $L=\lambda W$," Operations Research 9, 1961.

MOOR72 Moore, F.R. "Computation Model of a Closed Queuing Network with Exponential Servers," IBM J. Res. Develop. November, 1972.

MORR82 Morris, M.F., and Roth, P.F. Computer Performance Evalultion for Effective Analysis. Van Nostrand, New York, 1982.

MUNT78 Muntz, R.R. "Queueing Networks: A Critique of the State of the Art and Direction for the Future," ACM Computing Surveys, September 1978.

PARZ62 Parzen, E. Stochastic Processes. Elsevier North Holland, New York, 1962.

REIS75 Reiser, M., and Kobayashi, H. "Queueing Networks with Multiple Closed Chains: Theory and Comutational Algorithms," IBM J. Res. Develop., May, 1975.

REIS80 Reiser, M., and Lavenberg S.S. "Mean-Value Analysis of Closed Multichain Queueing Networks," J. ACM, April, 1980.

REIS82 Reiser, M. "Performance Evaluation of Data Communication Systems," IEEE Proceedings, February, 1982.

ROSS70 Ross, S.M. Applied Probability Models with Optimization Applications. Holden-Day, San Fancisco, 1970.

ROSS80 Ross, S.M. Introduction to Probability Models, second edition. Academic Press, New York 1980.

SAAT61 Saaty, T.L. Elements of Queueing Theory with Applications. McGraw-Hill, New York, 1961.

SAUE81 Sauer, C.H. "Approximate Solution of Queueing Networks with Simultaneous Resource Prossession," IBM J. Res. Develop., November, 1981.

SAUE81 Sauer, C.H., and Chandy, K.M. Computer Systems Performance Modeling. Prentice-Hall, New Jersey, 1981.

SAUE82 Sauer C.H. "Numerical Solutions of Some Multiple Chain Queueing Networks," Performance Evaluation Review 10, 4 (Winter 1981-1982).

SAUE83 Sauer C.H. "Computational Algorithms for State-Dependent Queueing Networks," ACM Transactions on Computer Systems, February, 1983.

SCHW87 Schwartz, M., Telecommunications Networks: Protocols, Modeling and Analysis. Addisson-Wessely, 1987.

STUC85 Stuck, B.W., and Authurs, E. A Computer and Communications Network Performance Analysis Primer. Prentice-Hall, New Jersey, 1985.

TAKA62 Takacs, L. Introduction to the Theory of Queues. Oxford University Press, 1962.

THOM69 Thomasian, A.J. The Structure of Probability Theory with Applications. McGraw-Hill, New York, 1969.

WHIT68 Whittle, P. "Equilibrium Distribution for an Open Migration Process," J. Appl. Prob. 5, 1968.

END

9-87

DTIC